J Forensic Sci, November 2003, Vol. 48, No. 6
Paper ID JFS2003029_486
Available online at: www.astm.org

Robert G. Cowell,[1] *Ph.D. and Petter Mostad,*[2] *Ph.D.*

A Clustering Algorithm Using DNA Marker Information for Sub-Pedigree Reconstruction*

ABSTRACT: In a mass disaster scenario in which many people are dead, it may be that small family groups are among the dead, and investigators may need to identify such groups, e.g., to return bodies to living relatives for burial. We consider the problem of identifying small groups of closely related people within a large group of people through the use of DNA marker information. We propose a likelihood-ratio–based distance measure of the relatedness between pairs of individuals and use an estimate of this measure as a means of clustering related people into groups. We show the effectiveness of our approach on real examples and through simulations, which suggest that the method is quite reliable for identifying very close relationships. We discuss the use of our clustering algorithm in a two-stage pedigree reconstruction procedure and suggest directions in which the analysis could be extended. Applications include the identification of family groups among bodies found in mass graves and identification of family groups in animal populations.

KEYWORDS: forensic science, mass graves, mass disasters, DNA markers, pedigree reconstruction

DNA markers are now routinely used to determine the correct familial relationship between persons or animals. In many applications, e.g., paternity cases, only two (or a very short list) of possible familial relationships, or *pedigrees,* are considered. But the technique can also be used in complex situations where there are a large number of possibilities, e.g., identifying victims of an aircraft disaster or a terrorist attack or identifying the bodies in a mass grave that requires the consideration of a large number of explanatory hypotheses.

In several of these applications, e.g., aircraft disasters, information from passenger lists, from relatives, and other nongenetic evidence will be combined to create a list of missing persons, and DNA evidence may be used to match the bodies to this list. In other applications, less nongenetic information may be available, and the more general problem of finding a correct familial relationship between a set of persons will come into focus. This would be the case with, for example, a mass grave with no corresponding missing person list. DNA testing has also been used to study relatedness in animal populations (1). In such contexts, less nongenetic information will usually be available.

This paper studies the problem of determining correct family relations in a large group of individuals, focusing on the case where no nongenetic information is available. We assume, however, that the true pedigree linking the individuals consists of many smaller sub-pedigrees with very distant relationships between the sub-pedigrees.

As we understand and model quite well the processes generating DNA marker data, we can formulate the problem as finding the pedigree or pedigrees with the highest likelihood (or the highest posterior probability) explaining the data. However, this "pedigree-reconstruction" problem will usually be intractable because of the complexity of the problem. The set of possible pedigrees connecting a set of observed people (i.e., people from whom DNA measurements have been taken) is infinite if one allows an unlimited number of unobserved people to enter into the pedigree. In general, at the present time, one must limit the number of unobserved people allowed in the description of the pedigree to make the problem tractable. This is the approach implemented in the computer program FAMILIAS (2,3). But beyond about five observed people and one or two unobserved people, the number of possible pedigrees exceeds what can be handled by the program, unless there are specific restrictions on the possible relationships between the persons. Thus, the general problem becomes intractable.

In several of the scenarios described above, we may exploit the (prior) information that most pairs of people will be unrelated to each other and that the pedigree sought will consist of many small disconnected sub-pedigrees of relatives. Then, one possibly efficient approach to the pedigree reconstruction is the following two-step *divide-and-conquer* algorithm: first divide the people into a sufficiently large number of small clusters such that people within a cluster are (strongly suspected to be) related, and people from distinct clusters are (strongly suspected to be) unrelated to each other. Then, provided the clusters are small enough, current software such as FAMILIAS could be used to reconstruct sub-pedigrees from them. The union of such cluster pedigrees will be the full pedigree. This is illustrated in Fig. 1.

In this paper we propose a measure of estimating the relatedness of two individuals and an algorithm that uses this measure to cluster individuals, that is, to perform the *divide* stage of the divide-and-conquer algorithm into possibly related familial groups of manageable size. We do not address the detailed sub-pedigree reconstruction of the small clusters, assuming that this is to be done using currently available software such as FAMILIAS (hence we do not consider complications arising from incest, population substructure, kinship. etc.). We shall also assume that other forensic information (such as dental records) or other detailed prior infor-

[1] Faculty of Actuarial Science and Statistics, Cass Business School, City of London, UK.
[2] Department of Math Statistics, Chalmers, Göteborg, Sweden.
* Supported in part by the Leverhulme trust.
Received 31 Jan. 2003; and in revised form 17 June 2003; accepted 18 June 2003; published 27 Aug. 2003.

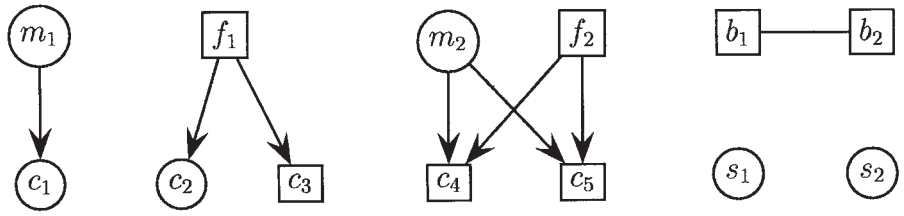

FIG. 1—*A pedigree of 13 observed people, consisting of six females (circles) and seven males (rectangles): female m_1 is the mother of c_1; the two females c_2 and c_3 are half-siblings sharing a common father f_1; the males c_4 and c_5 are full siblings by m_2 and f_2; the two males b_1 and b_2 are also full siblings, but their parents are not observed; the two females s_1 and s_2 are not related to each other nor to any of the other eleven people. The pedigree consists of the six connected components: $\{m_1,c_1\}$, $\{f_1,c_2,c_3\}$, $\{f_2,m_2,c_4,c_5\}$, $\{b_1,b_2\}$, $\{s_1\}$, and $\{s_2\}$. By correctly clustering the people into these six components, detailed pedigree relationships within each cluster could be found by current software. We are not aware of any software that could examine all possible pedigrees on 13 people (even without introducing unobserved people) because of the complexity arising from the large numbers of pedigrees to be considered.*

mation (such as the sizes and natures of the family groups) is unavailable to assist in the identification process.

The outline of the remainder of this paper is as follows. We begin by discussing the desiderata of a clustering procedure. We then define a distance measure appropriate in this context on which the clustering can be based and study its properties. We then propose a clustering algorithm, illustrating its use in some examples. Finally, we investigate the performance of the clustering algorithm in a simulation study.

Clustering Errors

In order to minimize the computational burden of the *conquer* stage, it is desirable to create as many clusters as possible, minimizing the size of each cluster as far as possible. Let *G* denote the set of people for which a pedigree reconstruction is desired. Now, for any two people, *A* and *B*, in the Set *G*, we have two possible hypotheses to entertain: (i) H_r, the two people are related, and (ii) H_u, the two people are unrelated. A clustering algorithm can make two possible decisions: allocate *A* and *B* to the same cluster, or allocate *A* and *B* to distinct clusters. Thus, there are two types of errors that a clustering algorithm may make that we call:

Type I Error: Allocate *A* and *B* to distinct clusters when they are in fact related.
Type II Error: Allocate *A* and *B* to the same cluster when they are in fact unrelated.

Ideally the clustering algorithm would be perfect and make neither type of error; however, this is unlikely to be the case. The best we can hope for is that the algorithm will try and minimize these errors. Of the two types of error, the Type II Error is not too bad in that, if two people who are not related are placed in the same cluster, their unrelatedness could yet be revealed in the full sub-pedigree reconstruction of the cluster. Nevertheless, it is desirable to minimize this type of error so that the sub-pedigree reconstructions can proceed as efficiently as possible.

In contrast, the Type I Error is bad, because if two related people are put into separate clusters during the divide stage, there is no way to recover from this error in the subsequent conquer stage. The extent to which our method makes this error indicates the extent to which it is an approximation compared to a full likelihood-based analysis of the data. Thus, we should aim that the clustering algorithm will make this type of error rarely, if at all.

The clusters generated by a clustering algorithm could either form a partition of the set of people *G* or they could form a covering set, that is, two distinct clusters may have people in common, with the union of the clusters giving the Set *G*. If the clusters form a partition, then clearly this will help to minimize the computational burden of finding the sub-pedigree of each cluster. However, this will tend to increase the possibility of both types of errors. For example, suppose that three people are related by being two unrelated parents and their common child. Then, if the clustering algorithm allocated the parents to distinct clusters and the clusters were to form a partition, the child would only be able to be in one of these two clusters. However, if the clusters formed a covering set, the child could be allocated to both clusters. After construction of the sub-pedigrees in each cluster, it would then be noticed that the child occurs in two of the sub-pedigrees, and one could consider a further stage in the pedigree reconstruction that looks at merging such overlapping sub-pedigrees. In this paper, we focus on generating a partition of *G* and so do not consider such merging strategies.

A Pair-Wise Genetic Distance

Definition of the Distance Measure

We now introduce our "distance measure" between two people based on their DNA profile. For more on ideas based on genes being identical by descent (IBD), see, for example, Ref 4. Note the following important fact: A pedigree influences the probability distribution of the autosomal DNA profiles of the two persons only through the "IBD constellation" it induces. We define an IBD constellation as a probability distribution on the set of IBD partitions of the four alleles of the two persons in one allele system. We define an IBD partition as a determination of which alleles (from each unordered pair) are IBD.

For example, the pedigree of a parent-child relation induces the IBD constellation that with Probability 1 we have the IBD partition where one allele from each person is IBD. A half-sibling pedigree induces the following IBD constellation: With Probability $\frac{1}{2}$, none of the alleles are IBD, and with Probability $\frac{1}{2}$, one allele from each person is IBD. For full siblings, we get that with Probability $\frac{1}{4}$ as none of the alleles are IBD, with Probability $\frac{1}{2}$, one allele from each person is IBD, and with Probability $\frac{1}{4}$, there are two pairs of IBD alleles, each pair with one allele from each person.

Although different pedigrees can induce an infinite number of possible IBD constellations, the most usual pedigrees linking persons *A* and *B* are: (i) *A* and *B* are parent and child (with either being the parent); (ii) *A* and *B* are full siblings or descendants from full siblings; and (iii) *A* and *B* are half siblings or descendants from

half siblings. Even if nonincestuous pedigrees can induce IBD constellations different from those induced by the pedigrees above (as incestuous ones also clearly can), such pedigrees are probably quite rare and not very important to consider in our two-step procedure, where the first step serves only to sort people into groups.

So what are the induced IBD constellations of these standard pedigrees? We have mentioned the one for parent and child and the one for full siblings. One may see that if one adds k_1 and k_2 successive descendant generations to each of these siblings, the induced IBD constellation has a probability of $(\frac{1}{2})^{k_1+k_2}$ that a pair of alleles is IBD (one allele from each person), while with Probability $1 - (\frac{1}{2})^{k_1+k_2}$, no alleles are IBD. Similarly, adding k_1 and k_2 successive descendant generations to each of two half-siblings, the induced IBD constellation has a probability of $(\frac{1}{2})^{k_1+k_2+1}$ that a pair of alleles is IBD, while with Probability $1 - (\frac{1}{2})^{k_1+k_2+1}$, no alleles are IBD. Based on this, it seems reasonable to make the following definition: *A* and *B* are said to have *distance i*, with *i* a positive integer, if the induced IBD constellation of the pedigree relating them has Probability $(\frac{1}{2})^{i-1}$ that a pair of alleles are IBD (one allele from each person), while with Probability $1 - (\frac{1}{2})^{i-1}$, no alleles are IBD. Note that the distance between a parent and its child is unity. We say that *A* and *B* have distance 1.5 if they are full siblings. We do not define the distance between persons related by pedigrees inducing other IBD constellations than these.

As mentioned, given the IBD constellation, we can compute the probability of observing given DNA data for the two persons: we simply condition on the different IBD partitions it can produce. Assume we have an allele system with alleles *a,b,c,*… and frequencies $p_a,p_b,p_c,\ldots$. Then, there are essentially five different types of DNA observations we can make for the two persons. (Note that the observations for the two alleles of each person are unordered and that we also need not be concerned with which of the two persons has which observation.) Their probabilities, given the IBD partition, are listed in Table 1.

It is useful to transform this table into one listing the likelihood ratio between the given IBD partition and the IBD partition where no alleles are IBD for each possible type of observation. This is done in Table 2. Given this information, we can now set up a table for the likelihood ratio of different distances between the two persons versus no relation between them (i.e., infinite distance), given DNA observations in one allele system. For example, if the observations are aa and aa, the likelihood ratio between distance 1.5 and infinite distance can be computed by considering the three possible IBD partitions: With Probability $\frac{1}{4}$, there are two pairs of IBD alleles, and the likelihood ratio is $1/p_a^2$; with Probability $\frac{1}{2}$, one allele from each person is IBD, and the likelihood ratio is $1/p_a$; and with Probability $\frac{1}{4}$, there are no IBD alleles, and the likelihood ratio is 1. Multiplying and summing gives the entry of the top right hand corner of Table 3; the rest of the table is derived similarly.

Let us assume we have two persons, *A* and *B*, and have observed data for them in *m* different loci. Define a_j, $j = 1,\ldots, m$ by

$$a_j = \left\{ \begin{array}{c} \frac{1}{p_a} - 1 \\ \frac{1}{2p_a} - 1 \\ \frac{p_a + p_b}{4p_ap_b} - 1 \\ \frac{1}{4p_a} - 1 \\ -1 \end{array} \right\} \text{when observations are} \left\{ \begin{array}{c} \text{aa, aa} \\ \text{aa, ab} \\ \text{ab, ab} \\ \text{ab, ac} \\ \text{no common alleles} \end{array} \right\}.$$

Then, assuming independence between allele systems, the total likelihood ratio for *A* and *B* having positive integer distance *i*

TABLE 1—*The probability of observing different types of data for two (unordered) persons, given the underlying IBD partition.*

Cases	None IBD	One Pair IBD	Two Pairs IBD
aa, aa	p_a^4	p_a^3	p_a^2
aa, ab	$4p_a^3p_b$	$2p_a^2p_b$	0
ab, ab	$4p_a^2p_b^2$	$p_ap_b(p_a + p_b)$	$2p_ap_b$
ab, ac	$8p_a^2p_bp_c$	$2p_ap_bp_c$	0
No alleles in common	(depends upon genotypes)	0	0

TABLE 2—*The likelihood ratio of different IBD partitions versus the partition where no alleles are IBD for different types of data for two (unordered) persons.*

Cases	One Pair IBD	Two Pairs IBD
aa, aa	$1/p_a$	$1/p_a^2$
aa, ab	$1/2p_a$	0
ab, ab	$(p_a + p_b) / 4p_ap_b$	$1/2p_ap_b$
ab, ac	$1/4p_a$	0
No alleles in common	0	0

TABLE 3—*The likelihood ratios of different distances (versus infinite distance) for different types of data for two (unordered) persons.*

Cases	Distance *i*	Distance 1.5
aa, aa	$\left(\frac{1}{2}\right)^{i-1}\frac{1}{p_a} + 1 - \left(\frac{1}{2}\right)^{i-1}$	$\frac{1}{4}\frac{1}{p_a^2} + \frac{1}{2}\frac{1}{p_a} + \frac{1}{4}$
aa, ab	$\left(\frac{1}{2}\right)^{i-1}\frac{1}{2p_a} + 1 - \left(\frac{1}{2}\right)^{i-1}$	$\frac{1}{2}\frac{1}{2p_a} + \frac{1}{4}$
ab, ab	$\left(\frac{1}{2}\right)^{i-1}\frac{p_a + p_b}{4p_ap_b} + 1 - \left(\frac{1}{2}\right)^{i-1}$	$\frac{1}{4}\frac{1}{2p_ap_b} + \frac{1}{2}\frac{p_a + p_b}{4p_ap_b} + \frac{1}{4}$
ab, ac	$\left(\frac{1}{2}\right)^{i-1}\frac{1}{4p_a} + 1 - \left(\frac{1}{2}\right)^{i-1}$	$\frac{1}{2}\frac{1}{4p_a} + \frac{1}{4}$
No alleles in common	$1 - \left(\frac{1}{2}\right)^{i-1}$	$\frac{1}{4}$

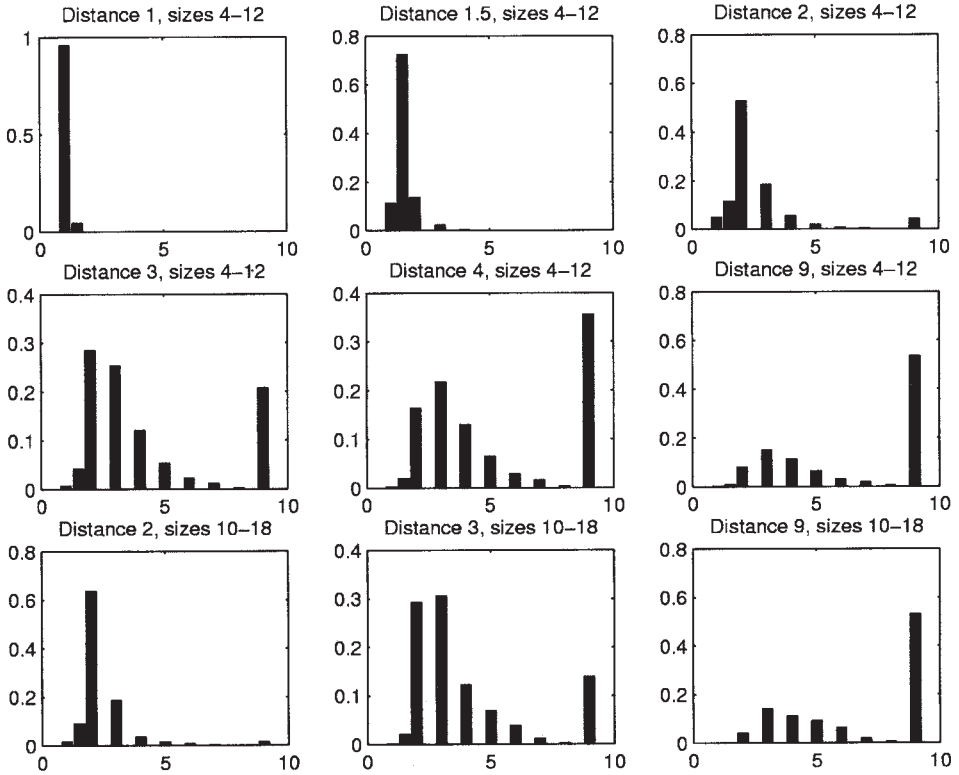

FIG. 2—*The figure shows the probability distributions of estimated distances for different true distances between two persons. The true distances are (from left to right and then top to bottom): 1, 1.5, 2, 3, 4, ∞, 2, 3, and ∞. In the plots, Distance 9 represents infinite distance. In the first six plots, nine systems with four through 12 alleles in each have been used, in the last three plots, nine systems with ten through 18 alleles were used.*

versus having infinite distance is

$$f(i) = \prod_{j=1}^{m} (2^{1-i} a_j + 1).$$

Defining

$$g(x) = \log\left(f\left(1 - \frac{\log(x)}{\log(2)}\right)\right) = \sum_{j=1}^{m} \log(x a_j + 1),$$

it follows that the second derivative of g is negative. Hence f is either monotonically increasing, monotonically decreasing, or reaches a unique maximum (for continuous i). Together with the fact that $\lim_{i\to\infty} f(i) = 1$, this makes it easy to maximize f.

The (smallest) positive integer i that maximizes f we shall call the *estimated distance* between A and B (unless the likelihood ratio for distance 1.5 is greater than this maximum; then we say that the estimated distance is 1.5). Note that it may well be infinite, corresponding to the two persons being unrelated. In the next section we illustrate the behavior of the distance measure.

Our definition of genetic distance is closely related to the kinship coefficient (or co-ancestry coefficient) between two people, defined as the probability that a randomly chosen allele from one of them will be IDB with a randomly chosen allele from the other (see, for example, Ref 5). Kinship may be used as an adjustment for unspecified relatedness when computing the probability of observed data given a pedigree. However, we can also think of it as a measure of genetic distance between two people. Two people with an integer distance i will have a kinship coefficient of 2^{-i-1}. However, full siblings, with a distance of 1.5 will have a kinship coefficient of 2^{-2}, the same as two people with Distance 1. It is clear that the kinship coefficient does not specify the IBD constellation, so we cannot compute the likelihood of a pedigree connecting two people given their kinship coefficient. However, if we assume that only the trivial IBD partition and the IBD partition with one pair (one allele from each person) are possible, then the kinship coefficient does determine an IBD constellation, and we can estimate the true kinship between two persons given DNA data in exactly the same way as we estimate distances. The results will also be the same, except that with our measure we also consider the possibility of a distance of 1.5.

Computations in Simple Cases

Having defined a distance between two persons and a way to estimate this distance, it is natural to ask how well the estimation works. In the most general cases, simulation is probably the best way to study this. However, if we make the simplification of only considering allele systems where all alleles have the same frequency, we can actually make explicit computations.

Assume we consider m such systems, with $n_1,\ldots, n_m$ alleles in each. With a given distance between two persons, we can for each system compute the probability of the persons having data of each of the five different types. We thus get the probability for each of the 5^m different total data types we can observe. For each such data type we can compute the maximum likelihood estimate for the distance as described above. In short, we get a probability distribution of different estimated distances.

As an example, Fig. 2 plots the distribution of the estimated distance for two persons of actual distance 1, 1.5, 2, 3, and ∞, when we have nine systems with four through twelve alleles in each system. It also plots the distribution for actual distances 1, 2, and ∞

when the systems have ten through 18 alleles. For the first set of systems, we see that Distances 1 and 1.5 are estimated quite well, and Distance 2 is estimated correctly in slightly more than half of all cases. But for Distances 3 and 4, estimation works increasingly badly, indicating that our method should not be used to estimate distances much above 2. Infinite distance is again estimated correctly slightly more than half the time. This means that almost half the time unrelated persons will be estimated to have some relation that can clearly cause problems in our two-step procedure; thus, it will be important to do the clustering in a careful manner. The last row of histograms indicates that using systems with more alleles tends to improve the estimation. This seems to be more important than using many systems.

A Clustering Algorithm

Recall that the aim of the clustering method is to subdivide the people in Set *G* into small clusters such that a group of people that are related are in the same cluster, while unrelated groups of people are assigned to different clusters. Any clustering algorithm is likely to be subject to the Type I and Type II errors described in the first section. Thus, ideally the clustering algorithm should aim to minimize the probability that a group of related people are split up and assigned to different clusters. However, a full solution to this optimization problem is likely to be intractable. To estimate the probability of a pair of people being related, we would need to set up a prior on how they might be related—or indeed unrelated—and then update the prior using the DNA data. A simplification to this full analysis would be to look at the likelihood ratio of the most likely relationship to that of being unrelated.

Here we do not pursue such a full or approximate probabilistic analysis. Instead we use the distances estimated from the likelihood ratios to find clusters in the following algorithm.

Partitioning Algorithm

From the members of Set *G* construct a graph as follows. Construct a graph with each person a node in the graph and initially without any edges in the graph connecting pairs of people. Select a fixed upper bound *d*. Then, for each distinct pair of members of the group connect them by an edge in the graph if their estimated distance is less than or equal to *d*. The clusters of interest are then the connected components of the graph. We can also describe this as doing hierarchical clustering with single linkage. The output could then be described in a dendrogram (6).

The choice of *d* will be crucial to the partitioning algorithm. In the limit of $d \to \infty$ we obtain a single cluster; hence, the probability of Type I errors will go to zero. As discussed in the second section, the results from Fig. 2 suggest that values of *d* greater than 2 will be subject to a great deal of Type II error. One approach in selecting *d* would be to run the partitioning algorithm with values of $d = 1$, $d = 1.5$, $d = 2$, etc., increasing *d* up to the largest value such that the largest cluster generated is still amenable to the software that would be used in the sub-pedigree analysis.

Examples

Example 1: Family with Two Children

Table 4 shows data on a family consisting of two full siblings and their parents. In Table 5 we show distance measures between each pair of individuals in the family evaluated using gene frequencies for an Italian population; the values in each row have been normalized to a maximum of unity. The table correctly identifies the mother and father as being unrelated (we took an upper limit of 10 for distances in our evaluations). All parent-child relationships are also correctly identified (Distance 1). The full-sibling relationship (Distance 1.5) is not picked out for the two children, but instead a Distance 2 is picked out, which if correct would indicate a grandparent-child or half-sibling relationship. However, the likelihood ratio is quite flat around the Distance 2, with Distance 1 being the next most likely. Thus the distance measure has picked out that the two children are closely related. The graph obtained using Distance $d = 1$ is shown in Fig. 3. The identification of this cluster

TABLE 4—*Data from a family consisting of two full siblings and their parents.*

Marker	Mother *m*	Father *f*	Child 1, c_1	Child 2, c_2
CFSPO1	10, 13	11, 12	11, 13	10, 11
D1S80	24	18, 26	24, 26	18, 24
D7S8	A	A, B	A, B	A
DQA1	1.1, 1.3	1.1	1.1, 1.3	1.1
GC	A, C	A, B	A, C	B, C
GYPA	A	A, B	A	A
HBGG	A, B	A	A, B	A
LDLR	A	A, B	A, B	A, B
TPOX	11	8, 9	9, 11	8, 11

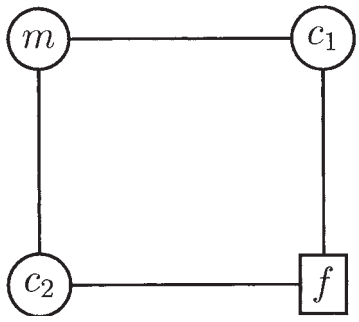

FIG. 3—*Graph obtained by connecting pairs of people in Table 4 if their estimated distance is equal to 1. Connecting people if their estimated distances were less than or equal to 2 would yield the same graph but with an extra edge connecting c_1 and c_2.*

TABLE 5—*Pair-wise normalized likelihood ratios for Example 1. Underlined values indicate the true distance (interpreting distance 10 as unrelated).*

	$i = 1$	$i = 1.5$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$
m:f	0.000	0.003	0.139	0.451	0.697	0.847	0.923	0.964	0.985	0.995	<u>1.000</u>
m:c_1	<u>1.000</u>	0.638	0.139	0.032	0.011	0.006	0.004	0.003	0.003	0.002	0.002
m:c_2	<u>1.000</u>	0.140	0.423	0.210	0.135	0.105	0.092	0.085	0.082	0.081	0.080
f:c_1	<u>1.000</u>	0.140	0.504	0.276	0.172	0.122	0.098	0.086	0.080	0.077	0.076
f:c_2	<u>1.000</u>	0.161	0.587	0.374	0.279	0.236	0.215	0.205	0.200	0.198	0.197
c_1:c_2	0.932	<u>0.239</u>	1.000	0.842	0.721	0.652	0.616	0.598	0.588	0.584	0.581

is sufficient for a fuller pedigree search using FAMILIAS, which in fact confirms that c_1 and c_2 are children of m and f.

Example 2: Four Children and One Parent

Our next example shows two full-siblings c_1 and c_2 and their mother m, and a further two full siblings, c_3 and c_4, by a different (unmeasured) mother, but who have the same (unmeasured) father as c_1 and c_2. The data are shown in Table 6. The likelihood ratios for various distances (normalized as in Table 5), are shown in Table 7. The results are reasonably encouraging. The mother-child relationships m:c_1 and m:c_2 appear strongly with peaks in the likelihood ratios at Distance 1 falling off very quickly with increasing distance. The full-sib relationship of c_1:c_2 also shows up strongly.

TABLE 6—*Data from a mother, m, her two full-sib children, c_1 and c_2, and two further full-sib children, c_3 and c_4, who share the same father as c_1 and c_2.*

Marker	m	c_1	c_2	c_3	c_4
CFS1PO	11, 12	11, 12	12	10, 12	11, 12
D13	10, 13	10	10, 11	12,13	11
D16	10, 13	10, 11	10, 11	12, 13	11, 12
D18	13, 19	17, 19	12, 13	15, 17	12, 14
D21	31.2, 33.2	30, 31.2	30, 31.2	28, 32.2	30
D3	15, 16	15, 16	15	15, 18	17, 18
D5	11, 12	12	11, 12	12, 13	12, 13
D7	10	7, 10	9, 10	9, 11	11
D8	9, 13	13, 14	13	12, 13	9, 13
HUMFGA	21, 22	22, 23	22, 23	23	21, 23
HUMTH01	7, 9.3	7	7	8, 9	7, 9
TPOX	12	8, 12	9, 12	9, 11	9, 12
vWA	16, 17	16	16, 17	16	16, 17

In contrast, the full-sib relationship of c_3:c_4 shows up quite weakly, although a half-sib relationship is indicated. Note that m:c_3 appear unrelated, although m:c_4 appear related at Distance 3. One of the four half-sib relationships (c_2:c_4) appears correctly at Distance 2, two appear at Distance 3 (c_1:c_3 and c_1:c_4), although the likelihoods are fairly flat around there, and one at Distance 4 (c_2:c_3), but again the likelihood is fairly flat. In summary, all children appear closely related, with a full-sib relationship appearing very strongly and two mother-child relationships also appearing strongly. The three graphs obtained using the clustering algorithms for upper bound values $d = 1$, $d = 1.5$, and $d = 2$ are shown in Fig. 4.

Example 3: The Romanov Data

Table 8 shows STR genotype data for the nine skeletons found in a shallow grave 20 miles from Ekaterinburg, Russia, and believed to be the remains of the Romanov family, some servants, and the family doctor, taken from Ref 7.

There are 36 different pairings of the nine skeletons, and we refrain from tabulating the likelihood ratios for every pair. Instead Table 9 shows a matrix of estimated distances evaluated for each pair of bodies. (We did not have the appropriate Russian population gene frequencies, so instead we used Italian population gene frequencies to construct this table.)

The graphs obtained by considering distances of $d = 1.5$ and $d = 2$ are shown in Fig. 5. We see immediately that the royal family members have been correctly grouped together and are separated from the other four people in the set of nine bodies. Including estimated distances of up to $d = 2$, we create a cluster consisting of the doctor and two of the servants; however, the servants appear unrelated to each other from Table 9.

The second step should be done with a likelihood-based procedure, analyzing all possible (reasonable) pedigrees connecting the

TABLE 7—*Pair-wise normalized likelihood ratios for Example 2. Underlined values indicate the true distance (interpreting distance 10 as unrelated).*

	$i = 1$	$i = 1.5$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	i = 8	$i = 9$	$i = 10$
m:c_1	<u>1.0000</u>	0.0113	0.0177	0.0009	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
m:c_2	<u>1.0000</u>	0.0114	0.0185	0.0011	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
m:c_3	0.0000	0.0002	0.0620	0.4052	0.7083	0.8653	0.9382	0.9720	0.9882	0.9961	<u>1.0000</u>
m:c_4	0.0000	0.0508	0.4915	1.0000	0.6403	0.3408	0.1981	0.1354	0.1072	0.0940	<u>0.0877</u>
c_1:c_2	0.0000	<u>1.0000</u>	0.0431	0.0054	0.0011	0.0004	0.0002	0.0001	0.0001	0.0001	0.0001
c_1:c_3	0.0000	0.0254	<u>0.8144</u>	1.0000	0.7166	0.5306	0.4374	0.3921	0.3700	0.3590	0.3536
c_1:c_4	0.0000	0.0192	<u>0.8559</u>	1.0000	0.6900	0.4920	0.3929	0.3447	0.3210	0.3093	0.3035
c_2:c_3	0.0000	0.0029	<u>0.4544</u>	0.9793	1.0000	0.9324	0.8822	0.8537	0.8387	0.8310	0.8271
c_2:c_4	0.0000	0.6234	<u>1.0000</u>	0.3284	0.1215	0.0615	0.0405	0.0320	0.0282	0.0264	0.0255
c_3:c_4	0.0000	<u>0.0339</u>	1.0000	0.9680	0.7244	0.5860	0.5177	0.4844	0.4679	0.4598	0.4557

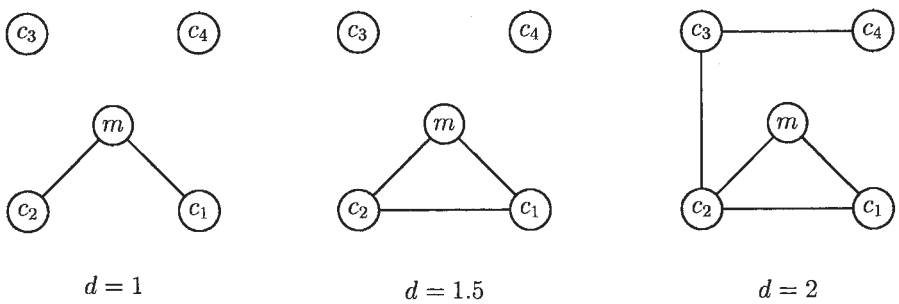

FIG. 4—*Graphs obtained using the partitioning algorithm for Example 2 of the text for the upper bound values of d = 1, d = 1.5, and d = 2 used as the criterion for connecting pairs of people using their estimated distances.*

TABLE 8—*Romanov STR data.*

Skeleton	HUMvWA/31	HUMTH01	HUMF12A1	HUMFES/FPS
1 (servant)	14, 20	9, 10	6, 16	10, 11
2 (doctor)	17, 17	6, 10	5, 7	10, 11
3 (child)	15, 16	8, 10	5, 7	12, 13
4 (Tsar)	15, 16	7, 10	7, 7	12, 12
5 (child)	15, 16	7, 8	3, 7	12, 13
6 (child)	15, 16	8, 10	3, 7	12, 13
7 (Tsarina)	15, 16	8, 8	3, 5	12, 13
8 (servant)	15, 17	6, 9	5, 7	8, 10
9 (servant)	16, 17	6, 6	6, 7	11, 12

TABLE 9—*Estimated distances for the Romanov STR data, based upon the four markers of Table 8.*

Skeleton	1 (s)	2 (d)	3 (c)	4 (Tsar)	5 (c)	6 (c)	7 (Tsarina)	8 (s)	9 (s)
1 (servant)	. . .	3	4	4	10	4	10	10	10
2 (doctor)	3	. . .	3	3	10	3	10	2	2
3 (child)	4	3	. . .	1	1.5	1.5	1.5	5	10
4 (Tsar)	4	3	1	. . .	1	1	3	6	3
5 (child)	10	10	1.5	1	. . .	1.5	1.5	5	10
6 (child)	4	3	1.5	1	1.5	. . .	1.5	10	10
7 (Tsarina)	10	10	1.5	3	1.5	1.5	. . .	10	10
8 (servant)	10	2	5	5	5	10	10	. . .	10
9 (servant)	10	2	10	3	10	10	10	10	. . .

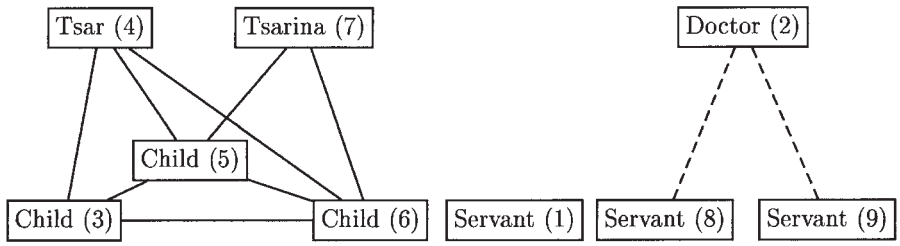

FIG. 5—*Graphs obtained by connecting pairs of people in the Romanov dataset using the estimated distances in Table 9. For $d \leq 1.5$ we obtain the graph with the full edges; for $d \leq 2$, the two extra broken edges between the doctor and two of the servants appears. Table 9 suggests that these two servants are unrelated, so taking into account that these estimates are based only on four markers, we conjecture that the broken edges represent Type II Errors.*

persons in each cluster. We illustrate how this can be done with the FAMILIAS program. When using this program, we need to assume knowledge of the sex of the involved persons (which is known in the Romanov case). The program can then generate all possible pedigrees containing the persons in the cluster (possibly adding additional persons to include more complex pedigrees) and find their likelihoods.

For the cluster of Size 5, the program finds 6720 pedigrees connecting them. Of these, the most likely pedigree is indeed the one where persons numbered 4 and 5 (the Tsar and Tsarina) are the parents of the three children, with a likelihood ratio of 2 to the next most likely pedigree. Excluding pedigrees where either of the persons 3, 5, or 6 are parents, we get 192 pedigrees, and the likelihood ratio increases to 116 between the previously found pedigree and the next most likely one, where Person 4 is not the father of Person 5.

For the possible cluster of Size 3, we add one extra female and one extra male, generating 1644 pedigrees. Of these, the most likely ones are those where there are parent-child relationships between the doctor and the two servants; these pedigrees are in fact more than five times as likely as the one where they are unrelated. Even if we based an analysis on other data that excludes the possibility that the three persons are in different generations, it is still more than twice as likely that the doctor is the half-brother of both servants as that they are all unrelated. This illustrates how difficult it is to determine correct familial relations based on only four marker systems.

Simulations

We now describe a simulation study that examined the sizes of clusters and the rates of occurrence of Type I and Type II clustering errors.

Our simulation is based upon the pedigree of 13 individuals, shown in Fig. 1. Given a set of k markers, and an upper distance radius d, an individual simulation consisted of drawing a random sample of 10,000 sets of genotypes for the 13 individuals, consistent with assuming both the Hardy-Weinberg equilibrium and also independence within and between markers. For each of the 10,000 sets of genotypes, a graph on the 13 individuals was constructed, with each node representing a distinct person, in which two distinct individuals were connected by an edge if and only if their estimated

distance was $\leq d$. From the graph the number of connected components and the size of the largest-connected component was found. It was also recorded for each pair of individuals whether or not they were in the same connected component. These values were accumulated over the 10,000 sets of genotypes. Independent simulations were performed for each of the 45 combinations of the three values (1, 1.5, 2) for d and all integer values of k from $k = 1$ up to $k = 15$. The gene frequencies used were estimates from a database of human markers. The markers used were, in order: PENTA_D, PENTA_E, CSF1PO, D13, D16, D18, D21, D3, D5, D7, D8, HUMFGA, HUMTH01, TPOX, and vWA. The simulations for k markers used the first k markers in this list.

Figures 6, 7, and 8 show the proportion of times that selected pairs of individuals were in the same cluster for distances $d = 1$, $d = 1.5$ and $d = 2$, respectively. From Fig. 6 we see that most of the time m_1 and c_1 (having true Distance 1) are in the same cluster, as are the full siblings (c_4 and c_5) and the half-siblings (c_2 and c_3). The reason the latter are generally in the same cluster, even though their true distances at 1.5 and 2, respectively, are both greater than the Distance Radius 1 used in constructing the cluster sets is that they tend to be connected to their parents at Distance 1. In contrast, the full siblings b_1 and b_2 (again true Distance 1.5) become less likely to be in the same cluster as the number of markers is increased, because neither of their parents were included in the clustering procedure. Similarly, the unrelated pair (m_1:s_1) tends to be in different clusters as the number of markers is increased, and the same is true for the unrelated pair (s_1:s_2). However, for a given number of markers, the proportion of times that (m_1:s_1) are in the same cluster tends to be higher than when (s_1:s_2) are in the same cluster; this is because most of the time m_1 is connected to c_1, and when this happens s_1 will be in the same cluster as m_1 if it is connected to *either* m_1 or c_1.

Figure 7 shows the effect of increasing the search radius to $d = 1.5$. The curves for the pairs (m_1:c_1), (c_2:c_3), and (c_4:c_5) all overlap with proportions equal to almost 1 for all values of k, thus correctly clustering them. The curve for the full siblings (b_1:b_2) increases smoothly from a proportion of approximately 87.5% at $k = 7$ to 94% at $k = 15$, in contrast to the downward trend in Fig. 6. The curves for the pairs (m_1:s_1) and (s_1:s_2) follow the same general shape as in Fig. 6, but their rate of decrease is much less.

The results in Fig. 8 are similar to those in Fig. 7. The curve for the pair (b_1:b_2) is now much closer to 1, having the lowest proportion of 94.8% at $k = 2$, reaching 98.9% at $k = 15$. The curves for (m_1:s_1) and (s_1:s_2) are now much higher than in Fig. 7 though still generally decreasing with increasing k.

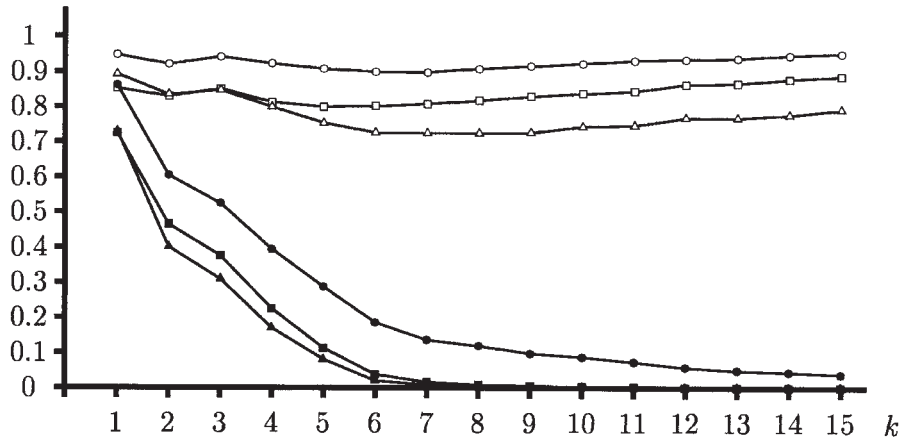

FIG. 6—*Fraction of times pairs are in the same cluster, for search radius d = 1. Open square (m_1:c_1), open triangle (c_2:c_3), open circle (c_4:c_5), filled circle (b_1:b_2), filled square (m_1:s_1), filled triangle (s_1:s_2). Horizontal axis gives the number k of markers.*

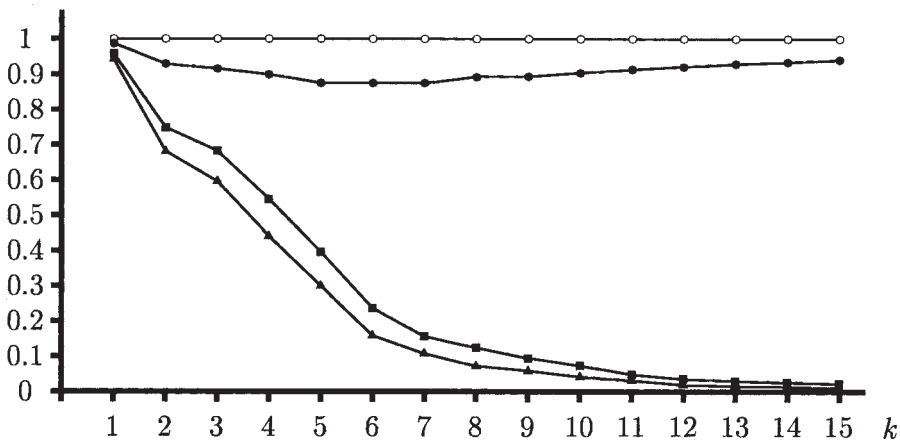

FIG. 7—*Fraction of times pairs are in the same cluster for search radius d = 1.5. Open circle (m_1:c_1, c_2:c_3, and c_4:c_5), filled circle (b_1:b_2), filled square (m_1:s_1), and filled triangle (s_1:s_2). Horizontal axis gives the number k of markers.*

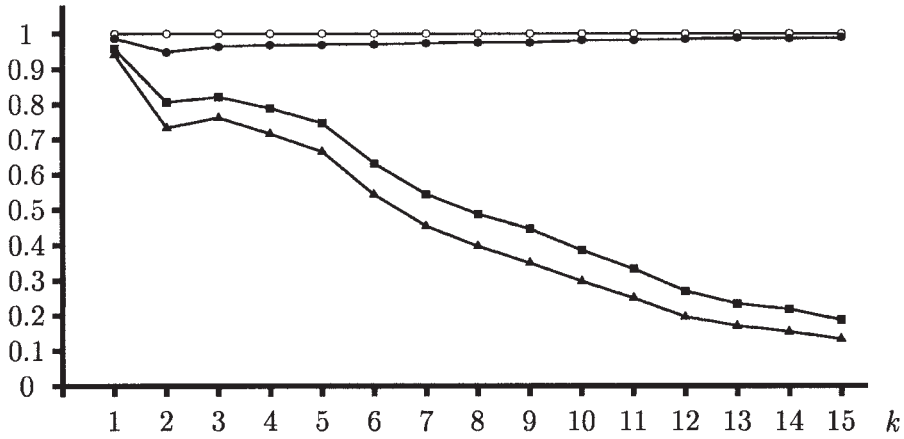

FIG. 8—*Fraction of times pairs are in the same cluster for search radius d = 2. Open circle (m_1:c_1, c_2:c_3, and c_4:c_5), filled circle (b_1:b_2), filled square (m_1:s_1), and filled triangle (s_1:s_2). Horizontal axis gives the number k of markers.*

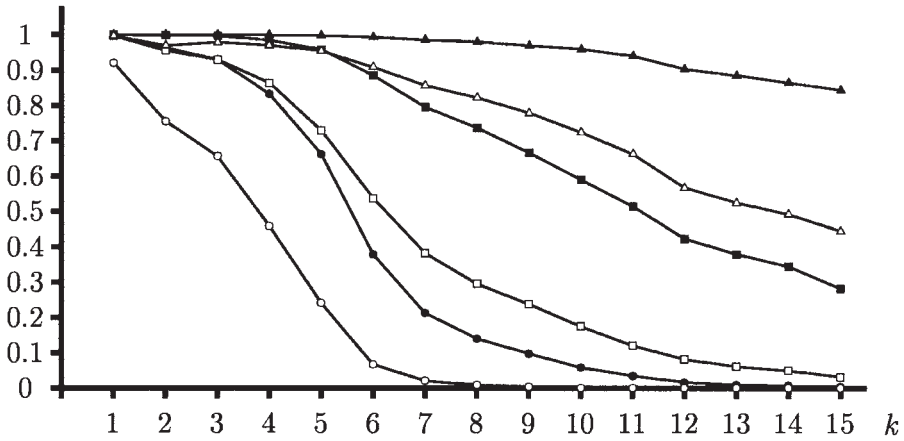

FIG. 9—*Fraction of clusters greater than or equal to size s generated during cluster searches of various radii d: open circles, s ≥ 8, d = 1; filled circles, s ≥ 5, d = 1; open squares, s ≥ 5, d = 1.5; filled squares, s ≥ 8, d = 1.5; open triangles, s ≥ 5, d = 2; filled triangles, s ≥ 8, d = 2.*

Figure 9 shows the dependence of the fraction of clusters greater than or equal to Sizes 5 or 8 for the three search radii d ∈ {1,1.5,2} upon the number of markers. The figure shows that more larger clusters are formed with greater search radii, but that their number decreases with increasing number of markers. The figure suggests that using a search radius of 2 or more is quite likely to produce sub-clusters that may be too large to analyze with current software.

Discussion

We have looked at the problem of identifying small family groups in a population of mostly unrelated individuals by using their DNA data for the purpose of performing the first step of a two-step procedure of identifying and reconstructing small family groups of individuals within a larger group. We have defined a distance measure between two individuals in a pedigree, shown how to estimate the distance between a pair of individuals, and have given an algorithm that generates a partition of individuals into groups using the estimated distances. We illustrated the effectiveness of our method on some examples and in a simulation study. These examples and simulations suggest that for identifying very close relationships, up to Distance 2 as we define it, the method is quite reliable with Type I error being quite rare, a desirable feature of a clustering algorithm.

There are several directions in which the analysis could be extended. One is to perform a Bayesian estimation procedure in which priors are put onto the various distances so that the whole posterior distribution of distances as a measure of the relationship between two persons may be found. This should lead to a better theoretical understanding of the Type I and Type II errors, with clustering based upon the posterior distributions. Another extension is to consider the sub-pedigree reconstructions of the family groups. There appear to be several ways this could be done. It is perhaps striking that in the Romanov example, almost all of the relationships seem to come out correctly for the members of the royal family by just considering the estimated pair-wise distances and using further prior information regarding the ages and sex of the five individuals without a further more detailed analysis. The same is true of the other examples. More generally, the estimated distances could be used in a pedigree search to restrict the pedigrees to look at. Thus, for example, if people *A* and *B* have an estimated distance of 1, one might consider only those pedigrees in which their distance is, say, either 1, 1.5, 2, or infinite. (The larger distances would be considered to allow for sampling error.) The search space would then become much smaller, thus increasing the computational efficiency and hence the complexity of the problems that could be tackled.

Acknowledgments

The authors would like to thank Marina Dobosz for providing the data and gene frequencies for the real examples.

References

1. Skaug HJ. Allele-sharing methods for estimation of population size. Biometrics 2001;57:750–6.
2. Egeland T, Mostad PF. Statistical genetics and genetical statistics: a forensic perspective. Scand J of Stat 2002;29:297–308.
3. Egeland T, Mostad PF, Mevåg B, Stenersen M. Beyond traditional paternity and identification cases. Selecting the most probable pedigree. Forensic Sci Int 2000;110:47–59.
4. Thompson EA, Heath SC. Estimation of conditional multilocus gene identity among relatives. In: Seillier-Moiseiwitsch, F, editor. Statistics in molecular biology and genetics, IMS lecture notes. Monograph series, Vol. 33;95–113. Hayward, CA: Institute of Mathematical Statistics, 1999.
5. Evett IW, Weir BS. Interpreting DNA evidence: statistical genetics for forensic scientists. Sinauer, 1998.
6. Johnson RA, Wichern DW. Applied multivariate statistical analysis. New York: Prentice Hall, 1998.
7. Gill P, Ivanov PL, Kimpton C, Piercy R, Benson, N, Tully G, et al. Identification of the remains of the Romanov family by DNA analysis. Nature Genetics, 1994;6:130–5.

Additional information and reprint requests:
Dr. R. G. Cowell
Faculty of Actuarial Science and Statistics
Cass Business School
City of London
106 Bunhill Row
London EC1V 8TZ U.K.
E-mail: rgc@city.ac.uk