

Symbolic Kinship Program

C. H. Brenner

Berkeley, California 94709

Manuscript received November 17, 1995
Accepted for publication October 10, 1996

ABSTRACT

This paper discusses a computerized algorithm to derive the formula for the likelihood ratio for a kinship problem with any arbitrarily defined relationships based on genetic evidence. The ordinary paternity case with the familiar likelihood formula $1/2q$ is the commonest example. More generally, any miscellaneous collection of people can be genetically tested to help settle some argument about how they are related, what one might call a "kinship" case. Examples that geneticists and DNA identification laboratories run into include sibship, incest, twin, inheritance, motherless, and corpse identification cases. The strength of the genetic evidence is always described by a likelihood ratio. The general method is described by which the computer program finds the formulas appropriate to these various situations. The benefits and the interest of the program are discussed using many examples, including analyses that have previously been published, some practical problems, and simple and useful rules for dealing with scenarios in which ancestral or fraternal types substitute for those of the alleged father.

A computer program, called the Kinship Program, calculates symbolic likelihood ratios, based on genetic evidence, for a general class of problems of which the ordinary paternity trio problem is the prototype. Examples include the following:

motherless case: Is this man the father of this child, based on genetic types from just the two of them?

incest case: Do the genetic types suggest that two people are doubly related?

sibling problems: Are two given people full siblings? half-siblings? unrelated?

inheritance problem: Are two people related as claimed?

twin problem: Are two siblings (whose parents are not tested) identical twins?

corpse identification: Is this corpse the same person who was reported missing by some family?

The inspiration for the Kinship Program was an earlier program developed by IHM (1975) and CONRADT (1983) that gives numerical answers to such problems. The novelty offered by the current program is that it produces explicit algebraic formulas. Naturally, once the formula is obtained a numeric answer can quickly and trivially be computed, so the formula is clearly as good as a number. In addition, the formula is a powerful tool that provides such advantages as verifiability, insight, and modeling.

In principle the formula may be an arbitrarily complicated rational function, a ratio of polynomials in the allele frequencies, and the time to derive it arbitrarily long depending on the complexity of the problem. However, the satisfying fact is that formula complexity

grows only slowly with the complexity of the problem. All practical problems that have arisen required only seconds on an ordinary desktop computer, and even more fanciful problems took at most a few minutes.

Background, paternity trios: As a foundation for the principles of analysis for the general case, we begin by reviewing the most familiar situation: the paternity test with mother, child, and an alleged father (whose paternity is to be decided) in a collection of genetic systems. Suppose that in some codominant system, such as a DNA test, the mother has genotype rp , child pq , and alleged father qs .

Let p, q, \dots represent the allelic frequencies corresponding to the alleles p, q, \dots , so that $2rp$ is the proportion of rp individuals in the population for example.¹ Then $2rp2qs$ is the proportion of rp, qs woman-man couples, and since $1/2 \cdot 1/2$ of such a couple's children are pq , the chance that a mother + child + father trio ("true trio") would have types rp, pq, qs is $2rp2qs \cdot 1/2 \cdot 1/2$. That is,

$$X = P\{\text{types as observed} | \text{true trio}\} = 2rp2qs \cdot 1/2 \cdot 1/2.$$

On the other hand, the chance that a "false trio" (woman + child + unrelated man) would have such types is

$$Y = P\{\text{types so observed} | \text{false trio}\} = 2rp2qs \cdot 1/2q,$$

so the likelihood ratio $X/Y = 1/2q$, which is well known (WALKER 1983).

The above analysis is idealized: the possibilities of mutation and of laboratory error are not included. Hardy-Weinberg equilibrium is not an essential assumption.

¹ Nonstandard typography is necessitated to distinguish numeric variables from genes.

Address for correspondence: Charles H. Brenner, Consulting in Forensic Mathematics, 2486 Hilgard Ave., Berkeley, CA 94709.
E-mail: cbrenner@ccnet.com

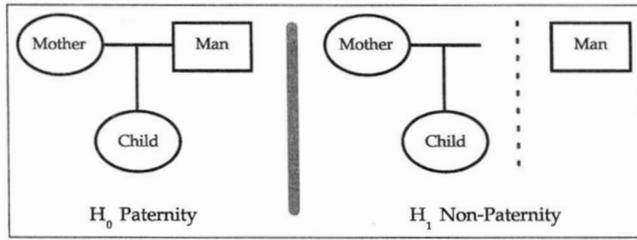


FIGURE 1.—The paternity problem posed as pedigrees.

tion for the simple trio situation above, although it will be assumed in the more general situations that follow. In the analysis the likelihoods X and Y could alternatively and equivalently have been defined with types of mother, man, and even mother's contribution as part of the conditions rather than included in the hypotheses, as here. The present point of view lends itself more readily to generalization however, so serves better as a model for what follows.

Pedigrees: To recapitulate, the ordinary paternity problem consists of comparing the two possible patterns of relatedness depicted in Figure 1. Some scientific evidence E , namely genetic types of the individuals involved, is determined. To assess the weight of the evidence E it is necessary and sufficient to evaluate the ratio X/Y , where $X = P\{E|\text{paternity}\}$ and $Y = P\{E|\text{non-paternity}\}$ (EDWARDS 1972).

Once the paternity problem is cast in this abstract way, clearly many generalizations of it can be solved in the same way. Instead of the specific relationships H_0 and H_1 of Figure 1, an arbitrary pair of "pedigrees," each of which may be a family tree or trees, may be compared.

Two grandparents case: The example shown in Figure 2 presents no new complications compared to the paternity case. One can write down by inspection that $X = P\{E|H_0\} = 2pr2qr2st \cdot \frac{1}{2} \cdot \frac{1}{4}$ where $2pr2qr2st$ represents the probability that the three independent ancestors would have types as shown; $\frac{1}{2}$ and $\frac{1}{4}$ represent the probabilities of p and q filtering down to the child from the maternal and paternal sides, respectively. Similarly, $Y = P\{E|H_1\} = 2pr2qr2st \cdot \frac{1}{2} \cdot q$, where q is the chance that a random sperm will have the allele q . Therefore $X/Y = \frac{1}{4q}$.

One grandparent case: The last case was simple in having only one possible origin for each of the child's alleles. The general situation is more complicated because there may be several combinations to consider.

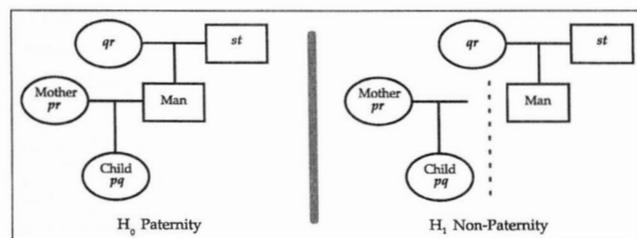


FIGURE 2.—Putative grandparents tested instead of man.

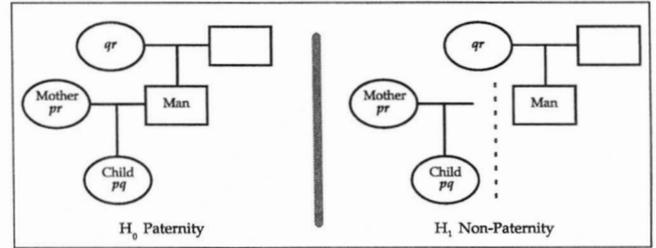


FIGURE 3.—One putative grandparent tested instead of man.

Figure 3 is like Figure 2 but the putative grandfather was not tested. Computation of X therefore requires considering two ways to transmit a q allele from the paternal lineage: $\frac{1}{4}$ that the putative grandmother passes her q to the child, plus $q \cdot \frac{1}{2}$ that the putative grandfather produces a q sperm, which subsequently passes to the child. Thus $X/Y = (\frac{1}{4} + q/2)/q$.

Paternal ancestors: Rewriting the preceding results as averages leads to an interesting observation. For the two grandparents case, note that $X/Y = (\frac{1}{2q} + 0)/2$; for one grandparent, $X/Y = (\frac{1}{2q} + 1)/2$. The various terms in the numerators can be recognized as paternity indices corresponding to the types of the various alleged grandparents, that is, interpret $\frac{1}{2q}$, 0, and 1 as the paternity indices when a qr person, an st person, or an untyped person, respectively, are evaluated as a possible father. Both as a preliminary to the general method and for its independent interest, it is worthwhile exploring the rule suggested by these examples.

Paternal ancestors rule: Suppose that a paternity problem involves no double relationships, and the genetic information available for a man consists of genetic types for some of his direct ancestors. Then his paternity index in each genetic system is the average of the paternity indices corresponding to the types of his parents. Moreover, this rule can be iterated through previous generations.

To prove this rule, let g, h, \dots be the various alleles at some locus. For an adult k define the vector of transmission probabilities $\tau_k(\cdot)$, where $\tau_k(g)$ is the probability for person k to pass allele g to an offspring given the types known for k and/or k 's ancestors. If there is no information about the genetic types, $\tau_k(g) = g$, the frequency in nature of the allele g . If the genotype is known, then each $\tau_k(g)$ is 0, $\frac{1}{2}$, or 1.

Using this notation the paternity index can be represented as follows. Let $\tau(\cdot)$ and $\tau_k(\cdot)$ be the transmission probability vectors corresponding to mother and to an alleged father k . Let C be the set of genotypes consistent with the phenotype of the child, so that $gh \in C$ means that the gh genotype gives the child's phenotype. We can say

$$X_k = \sum_{gh \in C} \tau(g)\tau_k(h), \tag{1}$$

and $Y = \sum_{gh \in C} \tau(g)h$. Then $L_k = X_k/Y$ is the paternity index for person k .

Now define $m(k)$ to be the mother of an untyped

person k , and $f(k)$ to be the father. Under many circumstances the condition

$$\tau_k(\cdot) = \frac{\tau_{f(k)}(\cdot) + \tau_{m(k)}(\cdot)}{2}$$

holds, which we will refer to as condition S. For example, S is always true for a codominant system. For a multilocus system or one with silent alleles, like Rh, S still holds provided k 's ancestors are unrelated, as does (1).

Assume condition S and assume that the mother and the man k are unrelated. Then

$$\begin{aligned} L_k &= X_k/Y = \left[\sum_{gh \in C} \tau(g)\tau_k(h) \right] / Y \\ &= \left[\sum_{gh \in C} \tau(g) (\tau_{f(k)}(h) + \tau_{m(k)}(h)) / 2 \right] / Y \\ &= \left[\sum_{gh \in C} \tau(g)\tau_{f(k)}(h) + \sum_{gh \in C} \tau(g)\tau_{m(k)}(h) \right] / (2Y) \\ &= \left[\left(\sum_{gh \in C} \tau(g)\tau_{f(k)}(h) \right) / Y \right. \\ &\quad \left. + \left(\sum_{gh \in C} \tau(g)\tau_{m(k)}(h) \right) / Y \right] / 2, \end{aligned}$$

i.e.,

$$L_k = [L_{m(k)} + L_{f(k)}] / 2, \quad (2)$$

which is the paternal ancestors rule.

Clearly the principle can be extended back through additional generations. For example, if the mother, $m(m(k))$, of $m(k)$, is typed in her stead, then

$$L_k = \frac{\frac{L_{m(m(k))} + 1}{2} + L_{f(k)}}{2}.$$

Here we have used that fact that, since $f(m(k))$ is untyped, $L_{f(m(k))} = 1$.

The condition S and therefore the paternal ancestors rule is rather general in its applicability. The mother may be typed or not; it works equally well in either case, so long as there exists a $\tau(g)$ giving an adequate description of the mother's genetic contribution. It is not limited to single locus codominant systems, but is often also valid for systems of haplotypes such as the Rh system. However, it does not apply to a compound system defined by simultaneously considering the various combinations from two independent loci. Therefore, Formula 2 has to be applied one locus at a time. Thus, the paternal ancestors rule applies to HLA-AB only to the extent that recombination is ignored.

Avuncular indices: It is worth comparing the simple situation of paternal ancestors just discussed with the more complex possibility that a sibling b of the alleged

father k is tested. That is, suppose we want to calculate the paternity index, L_b , of k by virtue of testing k 's brother b . Equivalently, we can say that we are testing the brother b for uncle-hood. Consequently the term *avuncular index* was proposed (MORRIS *et al.* 1988) for the likelihood ratio L_k in such a case.²

Avuncular index rule: In a system for which genetic types are available for child, alleged uncle b and perhaps the mother, but not for b 's brother k , and L_b is the paternity index for b , the avuncular index L_k is

$$L_k = (L_b + 1) / 2. \quad (3)$$

In particular, in systems that exclude b from paternity his avuncular index is $1/2$, and in systems where the paternity index is large, as a rule of thumb, the avuncular index is about half as large.

To derive (3), consider the genotypes of b 's parents. Label the parents' alleles \mathbf{pr} and \mathbf{qs} (not necessarily distinct), and say b received alleles \mathbf{pq} . Let $\tau_b(\cdot)$ and $\tau_k(\cdot)$ be conditional transmission probabilities vectors for b and k conditioned on the known phenotype of b . Now, $\tau_b(g) = P\{\mathbf{p}$ is a $g\} / 2 + P\{\mathbf{q}$ is a $g\} / 2$. Each of \mathbf{p} , \mathbf{q} , \mathbf{r} , or \mathbf{s} has probability $1/4$ to be transmitted by k , so

$$\begin{aligned} \tau_k(g) &= \frac{1}{4}(P\{\mathbf{p} \text{ is } g\} + P\{\mathbf{q} \text{ is } g\} + P\{\mathbf{r} \text{ is } g\} + P\{\mathbf{s} \text{ is } g\}) \\ &= \frac{1}{4}(2\tau_b(g) + 2\tau_0(g)), \end{aligned}$$

where $\tau_0(\cdot)$: $\tau_0(g) = g$ is the transmission vector of nature, which is to say of a random person. That is,

$$\tau_k(\cdot) = \frac{\tau_b(\cdot) + \tau_0(\cdot)}{2}.$$

Since $L_0 = 1$, (3) can now be derived in the same way as (2).

Cases can arise where the above rules can be combined. For example, suppose that a man is tested and has a paternity index of L in some system. Since his avuncular index is $(L + 1) / 2$, it follows by the paternal ancestors rule that the paternity index for his untested nephew is $((L + 1) / 2 + 1) / 2$.

General problem: However, it is certainly not always true that the transmission probabilities are an adequate substitute for genetic types. A situation wherein the method is inadequate is an untyped man with three children. Considering only his transmission probabilities seems to permit him to contribute a different allele to each one, but that is absurd. Hence at best the transmission probability approach is limited to analysis of single genetic events. Also, there does not appear to be a simple analogue of (3) corresponding to the case

² In Latin, unfortunately, *avunculus* means maternal uncle (RON GARNER, personal communication), but since there is no word in English derived from *patruus* (paternal uncle), avuncular index seems to withstand pedantic scrutiny.

of multiple typed siblings b, c, \dots of an alleged father k . This situation is discussed further below.

Consequently, to deal with arbitrary kinship problems, a dynamic combinatorial approach is necessary. In the worst case it is necessary to explore a tree whose nodes correspond to all the different possible combinations of genotypes that all the people may have. The next section describes an algorithm, called the Kinship Program, that explores the tree.

THE KINSHIP PROGRAM

The kernel of the program is a recursive subroutine that calculates X or Y , *i.e.*, calculates a likelihood $P\{E|H\}$, where H is a pedigree describing the hypothesized relationships and E is the known phenotypes. The recursion iterates per person, from oldest to youngest. At each recursive level the program loops through the possible genotypes for the person, and the probabilities corresponding to each of these subcases are added.

Notation: To make the explanation more explicit, further notation is needed.

People and relationships: Let the people be numbered $1, 2, \dots, k, \dots, n$, such that ancestors always precede descendants. In the following descriptions subscripts will always refer to people. Let m_k denote the mother of person k , where $m_k = 0$ if the mother is unspecified and is therefore to be taken as an unknown, untyped person; otherwise $0 < m_k < n$. In particular, $m_1 = 0$. In a similar way f_k denotes the father of k . Sometimes the alternate notation $m(k)$ or $f(k)$ will be used to avoid double subscripts. A pair of ordered sets $H = ((m_1, \dots), (f_1, \dots))$ defines a pedigree.

Alleles and allele frequencies: Let p, q, \dots be names for the discrete alleles observed in any person, and p, q, \dots be their respective frequencies of occurrence. All alleles never observed can be lumped together under the single name z , so that $p + q + \dots + z = 1$.

Ordered genotypes: It will be useful to think of ordered genotypes rather than just genotypes, so for purposes of describing the algorithm pq will mean that p and q are the maternally and paternally contributed alleles, respectively. As ordered genotypes, pq and qp each have frequency pq in the population.

Ordered genotype assignments: We will also need to consider assignments of ordered genotypes to the first k people, written $G = (G_1, G_2, \dots, G_k)$, where each $G_i = g_i h_i$, meaning that the i th person is considered to have inherited g_i maternally and h_i paternally. Let $\ell(G) = k$ be the length of G . Let $G + gh$ be $(G_1, G_2, \dots, G_k, gh)$, *i.e.*, like G with the addition of an $\ell(G) + 1$ st person with the ordered genotype gh .

Phenotypic data: For some people there are phenotypic typing results. Write $E = (E_1, E_2, \dots, E_n)$ for the set of phenotypes, where E_k is the phenotype for person k . Phenotypes can be written pq or p or the special symbol ϕ meaning that E_k is untyped. Since the relevant

fact about a phenotype is the genotypes that would manifest as that phenotype, we can think of pq as an abbreviation for $\{pq, qp\}$ and ϕ as abbreviating $\{pp, pq, \dots, qp, \dots, zz\}$. That way it makes sense to say $pq \in \phi$ or $pq \in \phi$. If $G_i \in E_i$ for $i = 1, \dots, \ell(G)$, then we will say G is compatible with E .

Recursive algorithm: Now consider the probability $P\{E|H, G\}$ defined as follows. G is an assignment of ordered genotypes to the first $\ell(G) = k$ people, and G is compatible with E . $P\{E|H, G\}$ means that given the pedigree H and assuming a genotype assignment G specified for the first k people, what is the probability to observe the evidence E ?

$$\text{If } k = n, \quad P\{E|H, G\} = 1.$$

$$\text{If } k < n, \quad P\{E|H, G\}$$

$$= \sum_{g,h} \tau_{m(k+1)}(g|G) \tau_{f(k+1)}(h|G) P\{E|H, G + gh\}, \quad (5)$$

where the sum is taken over all possible ordered genotypes $gh \in E_{k+1}$, and τ represents the transmission probabilities of the genes g and h from the mother and father of the $k+1$ st person. Specifically, if $j > 0$ then $\tau_j(g|G)$ is 0, $1/2$, or 1 according as G_j has 0, 1, or 2 copies of the gene g . And

$$\tau_0(g|G) = g, \quad (6)$$

meaning that an unknown parent contributes g according to the frequency of g in the world.

$$\text{Putting } k = 0, \quad P\{E|H, G\} = P\{E|H\},$$

which gives either X or Y depending on whether H is H_0 or H_1 . Then the likelihood ratio is X/Y .

Comments on the recursive formula: So the crux of the algorithm is recursive evaluation of the Formula 5.

Tree trimming: There are a few obvious steps to take to make the evaluation of (5) more efficient.

First, before recursion begins the phenotypes E are analyzed. Each person's phenotype E_k is written as a list of ordered genotypes. Then a backward pass is made through the relationships, from youngest person to oldest, and all ordered genotypes are removed that are incompatible with the relationships.

For example, if a child has phenotype pq and the mother is pr , then the list of possible child ordered genotypes is trimmed from $\{pq, qp\}$ to just $\{pq\}$, a potential 50% reduction in evaluation steps. In the other direction, if the father of the same child is untyped, instead of considering all possible ordered genotypes for the father it will be enough to consider those with a q . If for example the possible alleles are p, q, r , and z , the number of possibilities is thus reduced from 16 to seven. Notice that most of this reduction is recognized only because the genotypes are ordered (maternal contribution distinguished from paternal). Facilitation of tree-trimming is the main reason to deal with ordered genotypes.

It may happen that another backward pass will turn

up additional savings. Since the effort for each such pass is small and the benefit may be enormous, the trimming step is repeated until it has no further effect. The search reduction from this analysis can easily be 1000-fold.

Second, during the recursive evaluation of (5), the list of pairs gh , over which the summation occurs, is further limited by this rule: since there is no point in evaluating any $P\{E|H, G + gh\}$ that will have a coefficient of 0, only those g and h for which $\tau > 0$ are considered.

Pedigree factoring: Sometimes the pedigree H consists of disjoint pieces, that is, the set of people can be partitioned into two or more sets with no interrelationships. This is normally the case under the alternative hypothesis, where typically the child, mother, and her relatives form one set, while the man and his relatives are presumed to be an unrelated set.

The probability for the entire pedigree is then calculated as the product of the probabilities for each of the independent pieces. If the number of combinations to inspect is u and w for the two pieces, then the total computation time will be only proportional to $u + w$ instead of to uw . Since u and w may be large numbers, the improvement seems well worthwhile. In practice, though, since there are two scenarios to evaluate and only one of them can be factored, the likely benefit is nearly 50%—useful, but not spectacular.

Symbolic evaluation: Evaluation of (5) consists only of adding and multiplying. Since each coefficient τ is either a constant ($1/2$ or 1) or an allele frequency, each $P\{E|H\}$ is a polynomial in the allele frequencies. This simple observation suggests doing the evaluation symbolically, adding and multiplying polynomials where the allele frequencies are letters, instead of arithmetically. The symbolic operations of course take longer, but almost only by a constant factor.³ Moreover, that constant is only about five, a small penalty thanks to the fact that the multiplications are all of a particularly simple kind: a polynomial times a monomial. No cross products arise. And the benefits are considerable; the symbolic program is much more interesting. The formula gives more information and has many uses.

DISCUSSION

Examples: The examples in this section are all calculated by the Kinship Program. First are two typical practical problems.

Inheritance problem: Arthur and Beatrice have different (dead) mothers. Arthur claims to be the sole child and heir of his dead father, but Beatrice believes she had the same father. She hopes to quantify the evidence of common alleles between herself and Arthur to establish her right to share the inheritance.

³ Because the tree depth is not increased, only the time to visit each node. To be precise, the visit time is not strictly independent of the size of the polynomial, but grows very slowly.

Suppose that Beatrice is right. Half-siblings figure to share no alleles in up to half of genetic systems (if the systems are highly polymorphic). In a genetic system where they have no similar alleles, the evidence is modestly against relationship. The likelihood ratio is $1/2$. In most of the remaining systems the sharing pattern will be pr, ps and the likelihood ratio in favor of half-sibship is $1/2 + 1/(8p)$.

There is a subtle point to observe in applying these formulas. It is important to use a realistic estimate for the frequency p , not a conservative one as is the custom in paternity work. In a simple paternity case, paternity can eventually be proven by using sufficiently many markers even if the true evidentiary strength of each marker is squandered by using conservative allele frequency estimates. But accumulating the evidence in a case like this is like running up a down escalator. With every other system relentlessly chopping the likelihood ratio in half, the forward steps must be efficient lest there be no progress at all.

For example, suppose there are five systems with no shared alleles and five systems where an allele with frequency 0.05 is shared. Then the correct overall likelihood ratio is $(1/2)^5(1/2 + 5/2)^5 = (3/2)^5 \approx 8$, substantial evidence for Beatrice's point of view. But if the allele frequencies were "generously" estimated as 0.1, the overall likelihood ratio would be estimated as $(1/2)^5(1/2 + 5/4)^5 = (7/8)^5 \approx 0.5$, unfairly favoring Arthur.

Another half-sibling problem: Three men are either brothers or half-brothers. They all have a common mother, but there are two fathers. The middle man, Milton, is a full brother either with the eldest, Ebenezer, or with the youngest, Yancy, but Ebenezer and Yancy have different fathers. DNA testing is done on Ebenezer, Milton, and Yancy to decide who Milton's father was.

In three different loci, the sharing patterns for the three men are as follows:

	Ebenezer	Milton	Yancy
Locus 1	pq	qr	qs
Locus 2	pq	pr	s
Locus 3	pq	qs	pr

For the first locus the likelihood ratio is 1, since the pattern of genotypes is symmetrical about Milton.

At locus two, only Ebenezer and Milton share an allele. Superficially this suggests nominating them as the full-brother pair, but the likelihood ratio again is unity. A combinatorial argument shows why: the p must be maternal, for otherwise the mother would have passed three different alleles to the three sons. Consequently, there is no evidence either way about fathers.

Locus three suggests that Milton is more likely to be Ebenezer's brother, since they might share a paternal q allele. How much more likely? The answer from the

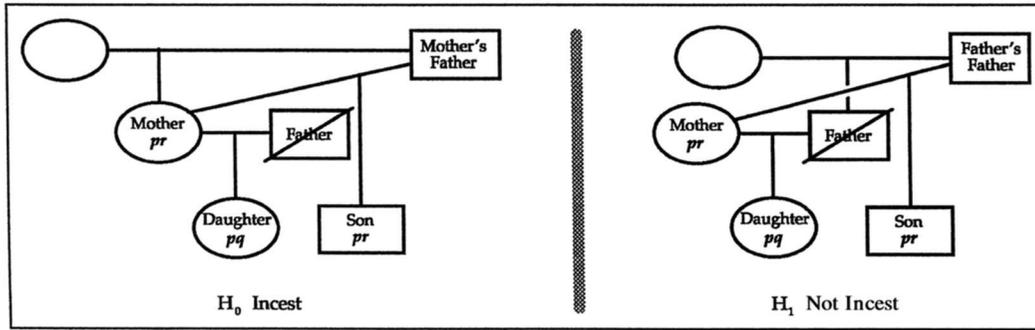


FIGURE 4.—An incest case. Son is Daughter's half brother. Son's father is one of Daughter's grandfathers. But which grandfather?

Kinship Program is $1 + 1/(2p + 2q)$. If p and q are both rare, the evidence is quite strong.

An incest case: Figure 4 shows a complicated incest case, the discussion of which will give a fair flavor of the workings of the program.

As illustrated by the diagram, Mother has two children. The father of the first child, Daughter, is the dead man called Father. The father of the second child, Son, is one or the other of Daughter's grandfathers. The problem is to decide which grandfather. Genetic types are available just for Mother, Son, and Daughter.

The two scenarios to be compared are described to the program in the following notation:

Daughter pq : Mother pr + Father

The $:$ separates the child's name from her parents. The parents are separated by $+$. Lowercase letters, e.g., pq , optionally following a name are the phenotype of that person.

Son pr : Mother + MothersFather / FathersFather

The $/$ separates two possibilities for a person. Son's father is MothersFather under one scenario and FathersFather under the other scenario.

Mother : ? + MothersFather

Mother's parents are here defined to be an unknown person (?) and MothersFather.

Father : ? + FathersFather

Father's parents are a different unknown person and FathersFather.

On initial analysis, there are four alleles to consider (p , q , r , and z). Therefore there are 16 possible ordered genotypes for each of the untyped people, Father, MothersFather, and FathersFather, and two possibilities for each of Mother, Daughter, and Son. That amounts to $\sim 65,000$ full length combinations for G to consider. Tree-trimming reduces this number to ~ 8000 , and in the end the number of evaluations of (5) is ~ 900 .

The variable z is undesirable because it has no meaning to the user. Besides, it is unnecessary. It can be eliminated using the relation $z = 1 - p - q - \dots$. Elimination of z produces a great simplification, reducing X from 20 terms to three and Y from 16 to two. It

is not clear why the disparity should be so great, nor even why the z -less form is the simpler, but typically it is so.

The final answer, after removing common factors from numerator and denominator, is $(2 + 2p + 2r)/(3p + 3r)$.

More avuncular indices: Armed with the general algorithm, we can delve further into the situation where one or more siblings of the alleged father are tested. Suppose as before that mother is p and child is pq . Table 1 shows the paternity index for a variety of combinations of genotypes and numbers of siblings. Of course there are patterns to the formulae, but apparently there is no simple rule to combine multiple uncles and aunts in the way that there is for ancestors.

The next two examples explore published results.

Twins: For twins to have the same genotype pq is evidence that they are monozygotic rather than dizygotic. Were the parents known to be pr and qs , the likelihood ratio would be exactly four. However, in the absence of typing the parents the possibility that one or both parents are homozygous reduces the likelihood ratio somewhat. The Kinship Program gives the formula $4/(1 + p + q + 2pq)$ when the parents are not typed and the twins are heterozygous pq . In a system for which the twins are homozygous qq , the likelihood ratio is $4/(1 + q)^2$. The latter formula especially is not difficult to verify by hand and is also given in VOGEL and MOTULSKY (1986, p. 671). AKANE *et al.* (1991), on the other hand, gave different formulas for these situations.

The question of determination of zygosity of twins occurs frequently. Most often the reason is that one twin has leukemia and needs a bone marrow transplant. If the genetic match is perfect, then the recipient can be spared the risk of immunosuppressants. Since there is little theoretical benefit in typing the parents and neither leukemia nor twinship is particularly rare, the problem probably arises dozens of times per year.

Silent alleles and paternity cases: Among many possible enhancements to the program, a quite easy one was the inclusion of silent alleles. From the point of view of the kernel program this meant only adding more possible ordered genotypes corresponding to the homozygous phenotypes: p is now $\{pp, po, op\}$ where o is the silent allele.

TABLE 1

Two singly infinite and one doubly infinite family of avuncular indices, where the paternal contribution is q

No. of pq siblings of alleged father	No. of qq siblings of alleged father				
	0	1	2	3	4
0	1	$\frac{1}{2q} + \frac{1}{2}$	$\frac{1}{2q} + \frac{1}{1+q}$	$\frac{1}{2q} + \frac{2}{1+3q}$	$\frac{1}{2q} + \frac{4}{1+7q}$
1	$(1/4q)(1+2q)$	$\frac{1}{2q} + \frac{1/2}{1+q}$	$\frac{1}{2q} + \frac{1}{1+3q}$	$\frac{1}{2q} + \frac{2}{1+7q}$	$\frac{1}{2q} + \frac{4}{1+15q}$
2	$\frac{(1/4q)(1+3q-2q^2(1+3p))}{1+p+q+2pq}$	$\frac{1}{2q} + \frac{1}{1+p+3q}$	$\frac{1}{2q} + \frac{2}{1+p+7q}$	$\frac{1}{2q} + \frac{4}{1+p+15q}$	$\frac{1}{2q} + \frac{8}{1+p+31q}$
3	$\frac{(1/4q)(1+5q-12q^2(1+5p))}{1+3p+3q+12pq}$	$\frac{1}{2q} + \frac{2}{1+3p+7q}$	$\frac{1}{2q} + \frac{4}{1+3p+15q}$	$\frac{1}{2q} + \frac{8}{1+3p+31q}$	$\frac{1}{2q} + \frac{16}{1+3p+63q}$

As an application, Table 2 covers all patterns of shared alleles, includes the case where the mother is not typed, and gives the formula both when silent alleles are considered possible and when not.

The last part of Table 2 covers the cases with an untyped mother. BRENNER (1993) lists the motherless formulas for codominant systems, and the computer confirms these simple formulas. CHAKRABORTY *et al.* (1994) give formulas for the motherless case and silent alleles. In the first five of six cases the Kinship Program agrees. In the last case (the so-called "indirect exclusion," a q child and r man) they gave $1/4o$, o being the frequency of silent alleles, not a plausible formula since it says that

the more rare is the silent allele, the more likely are father and child to have it. LUQUE and VALVERDE (1996) have also made this point.

Limitations: The program as described is general but does have some restrictions. The genetic systems must be codominant, or at most to have a silent allele. On the face of it this restriction could easily be removed: any autosomal genetic system can be set up by having a table containing genotype lists that correspond to each possible phenotype. Such a scheme would encompass Rh or MNSs, but not two loci of HLA unless recombination were ignored.

The present version presumes that all allele frequen-

TABLE 2

Likelihood ratios for various paternity situations

Child	Mother	Tested man	Likelihood ratio for paternity <i>vs.</i> nonpaternity	
			Codominant system	With silent allele o
q	pq	q	$1/q$	$1/q'$
q	p	q	impossible	$(q'/q'')/q$
pq	p or pr	q	$1/q$	$(q'/q'')/q$
q	q	q	$1/q$	$1/(q' - o^3/q'(q+3o))$
pq	p or pr	qr	$1/2q$	
q	p	qr	impossible	$1/2q$
q	pq	qr	$1/2q$	$1/2q'$
q	q	qr	$1/2q$	$1/2(q' - o^2/q'')$
pq	pq	pq	$1/(p+q)$	
pq	pq	q	$1/(p+q)$	$(q'/q'')/(p+q)$
pq	pq	qr	$1/(2p+2q)$	
q	pq	r	0	$o/q'r''$
q	q	r	0	$o/(q'r'' + oqr''/q')$
q	p	r	impossible	0
q	Mother not tested	q	$1/q$	$(1 + o(1 - o/q'')/q)/q''$
pq	Mother not tested	q	$1/2q$	$(q'/q'')/2q$
q	Mother not tested	qr	$1/2q$	$(q'/q'')/2q$
pq	Mother not tested	pq	$(p+q)/4pq$	
pq	Mother not tested	pr	$1/4q$	
q	Mother not tested	r	0	$o/q''r''$

If the possibility of an undetected allele changes the formula, the more general form is listed in the silent allele column, using abbreviations $q' = q + o$, $q'' = q + 2o$, $r'' = r + 2o$, where o is the frequency of the silent allele.

cies are from the same race. Allowing various races presents no difficulty in principle; it would affect only (6). Instead of a single symbol θ to denote an untyped ancestor, there need to be $\theta, \theta', \theta'',$ etc. for the various races. Then, additional symbols $g', g'',$ etc. would be introduced for allele frequencies of the extra races, where $\tau_{\theta'}(g) = g',$ etc.

A few more restrictions are incidental to the kernel of the program and are only limitations of the Kinship input language through which the user describes the scenarios. For example, it might be more convenient if the program allowed comparison of more than two scenarios at once, although this is no limitation in principle because the scenarios can always be compared pairwise. The input description language necessitates that both scenarios include the same set of children. To describe the mono-/dizygotic twin scenarios therefore requires an artifice or "programming trick."

Alleles must be discrete. With restriction fragment length polymorphism systems, where the reality is a collection of sizes none of which match exactly, the user has to decide which measurements represent identical alleles and which do not. In practice this is acceptable, but it would not be prohibitively hard to write a more general program that could deal with continuous allele measurements and measurement error, nor would such a program necessarily be very slow to run.

Conclusions: One motive for writing the Kinship Program is that working out these problems by hand is very prone to error, as is shown by published errors.⁴ The program is interesting and useful because it gives clear and correct answers. The benefits of this include the following.

If the result is to be presented in an adversarial setting (in court), the formula can be given as justification for the calculation. Since the formula could perhaps be doubted as well as a number this justification at first sounds a bit circular, but in practice it is very helpful to be provided with this intermediate result. Typically, the formula can confidently be verified by hand even if deriving it *de novo* would be very chancy.

The formulas can be instructive, surprising, and revealing. The idea that realistic rather than conservative allele frequencies are necessary for half-sibling (and many other) cases is one example that is apparent from consideration of the formulas but previously escaped attention.

Such rules as the simple general formulas for paternal ancestors and for uncles are more likely to be apparent given the relatively abstract point of view provided by symbolic likelihood ratios. The precise scope of these rules is hard to characterize. It depends in part on the complexity of the typing system. For example, consider

the allele transmission probabilities implied by knowledge of the parental types of an untyped alleged father. If the system is codominant, then also typing the alleged father's uncle would change nothing, whereas for a more complicated system, such as ABO or Rh, typing the uncle adds information. The complexity of the relationships also bears on the applicability of the rules. On its face Equation 1 depends on the independence between maternal and paternal transmission probabilities; nonetheless in practice rule (2) holds even for some examples where mother and alleged father are related (while failing for similar examples). Also, there is interplay between the complexity of the genetic system and the degree to which the paternal ancestor or avuncular rule are tolerant of such relationships.

The static "transmission probability" approach contrasts with the recursive algorithm (or "combinatorial approach") embodied by the Kinship Program. The latter approach is necessary in general, but the principle underlying the former, which is embodied in the rules (2) and (3), applies in parts even to problems to which it does not provide a complete solution. Especially (3) is conservative of formula complexity, in that it creates no new terms when applied to a ratio of polynomials. Thus the principle represents an effective reduction that is one reason that the formula for even a large problem may well be simple.

Thanks to M. MCGINNIS and J. THOMAS for providing interesting examples. I am especially indebted to B. WEIR for invaluable advice and encouragement in preparing this paper.

LITERATURE CITED

- AKANE, A., K. MATSUBARA, H. SHIONO, M. YAMADA and Y. NARAGOME, 1991 Diagnosis of twin zygosity by hypervariable RFLP markers. *Am. J. Med. Genet.* **41**: 96-98.
- BRENNER, C., 1993 A note on paternity computation in cases lacking a mother. *Transfusion* **33**: 51-54.
- CHAKRABORTY R., L. JIN and Y. ZHONG, 1994 Paternity evaluation in cases lacking a mother and nondetectable alleles. *Int. J. Leg. Med.* **107**: 127-131.
- CONRADT, J., 1983 Serostatistische Abstammungsbegutachtung: ein Algorithmus für Verwandtenfälle und das Daten- und Programmsystem PAPS (dissertation). Görich & Weiersgäuser, Marburg Germany.
- EDWARDS, A. W. F., 1972 *Likelihood, An Account of the Statistical Concept of Likelihood and Its Application to Scientific Inference*. Cambridge University Press, Cambridge.
- IHM, P., and K. HUMMEL, 1975 A method to calculate the plausibility of paternity using blood groups results of any relatives. *Z. Immunitätsforsch. Exper. Klin. Immunol.* **149**: 405-416.
- LUQUE, J. A., and J. L. VAVERDE, 1996 Paternity evaluation in cases lacking a mother and non-detectable alleles (letter). *Int. J. Leg. Med.* **108**: 229.
- MORRIS, J. W., R. A. GARBER, J. D'AUTREMONT and C. H. BRENNER, 1988 The avuncular index and the incest index. *Adv. Forensic Haemogenet.* **2**: 607-611.
- VOGEL F., and A. G. MOTULSKY, 1986 *Human Genetics: Problems & Approaches*, Ed. 2. Springer-Verlag, Berlin.
- WALKER, R. A., 1983 *Inclusion Probabilities in Paternity Testing*, Amer. Assoc. of Blood Banks, Arlington, VA.

Communicating editor: B. WEIR

⁴ Unedited computer output will gladly be supplied on request for any problem discussed in this paper.