Probability of Gene Identity by Descent: Computation and Applications

Alice S. Whittemore, Jerry Halpern

Biometrics, Volume 50, Issue 1 (Mar., 1994), 109-117.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199403%2950%3A1%3C109%3APOGIBD%3E2.0.CO%3B2-V

BIOMETRICS 50, 109–117
March 1994

Probability of Gene Identity by Descent: Computation and Applications

Alice S. Whittemore and Jerry Halpern

Department of Health Research and Policy, Stanford University School of Medicine,
Stanford, California 94305, U.S.A.

SUMMARY

Two genes at a given locus are identical by descent (IBD) if both have been inherited from a common ancestor. We present an algorithm for computing the probabilities of all IBD relationships among the genes of pedigree members. We show how to use these probabilities to calculate the probability of any combination of genotypes or phenotypes for the pedigree members. Applications to linkage analysis and genetic counseling are illustrated with examples. The algorithm also can be used to calculate the generalized kinship coefficients proposed by others.

1. Introduction

Two genes at a given locus are identical by descent (IBD) if both have been inherited from a common ancestor. For example, in a family without inbreeding, two sibs who have inherited the same gene from their father but two different genes from their mother have one gene (the paternal one) IBD. The concept of identity by descent was introduced by Cotterman (unpublished Ph.D. thesis, Ohio State University, 1940) and extended by many authors, including Malecot (1948), Li and Sacks (1954), Cockerham (1971), Jacquard (1972), Denniston (1974), Thompson (1974), and Karigl (1981, 1982). Basic to the concept is the idea of an IBD configuration of genes [also called an orbit (Thompson, 1974) or a condensed identity state (Karigl, 1982)] for n individuals that specifies which of the individuals' $2n$ genes are identical by descent. IBD configurations are used in analyzing how a trait is inherited within pedigrees, to calculate the probabilities of observed phenotypes for pedigree members. In linkage analysis, probabilities of IBD configurations give the distribution of marker genotypes among pedigree members. In genetic counseling, they are used to estimate the probability that a family member develops a certain condition or that a prospective child would have such a condition, given the occurrence of this condition in the family pedigree. In addition, they can be used to estimate allele frequencies in a population from the allele occurrence in related individuals (Boehnke, 1991).

Thompson (1974) derived general formulae for the number of IBD configurations possible among n individuals. Table 1 shows that the number of such configurations grows rapidly with n, even when attention is restricted to pedigrees without inbreeding. Not all configurations are equally likely, given the genealogical relationship of the individuals. Indeed, many of the possible configurations are inconsistent with their relationship, and so have probability zero. In Section 2 we present an algorithm for computing the probabilities of all IBD configurations among a set of pedigree members that are consistent with the relationship. Ethier and Hodge (1985) give formulae for such probabilities when the pedigree members are full siblings. The algorithm presented here is analogous to the peeling algorithm of Cannings, Thompson, and Skolnick (1978). The latter is a recursive scheme for calculating the probability of an observed set of phenotypes for the pedigree members. In contrast, here we calculate the probabilities of all IBD configurations for the members that are consistent with the pedigree. In Section 3 we show how to use these probabilities to calculate the probability of any combination of genotypes or phenotypes for the pedigree members, as well as the generalized kinship coefficients of Karigl (1982) and Weeks and Lange (1988).

Key words: Allele; Condensed identity coefficient; Gene-identity state; Generalized kinship coefficient; Identical-by-descent.

Table 1
The number of possible IBD *configurations among n noninbred individuals*

Individuals	Configurations
2	3
3	16
4	139
5	1,750
6	29,388
7	624,889
8	16,255,738
9	504,717,929
10	18,353,177,160
11	769,917,601,384
12	36,803,030,137,203

2. Computing IBD Configuration Probabilities

We begin by defining an IBD configuration for the $2n$ genes of an ordered set of n individuals in a pedigree. We assign to the set a vector $s = (s_{11}, s_{12}, \ldots, s_{n1}, s_{n2})$ of gene labels in the following way. First we order each individual's genes by specifying that the paternal gene precedes the maternal one. This arbitrary ordering, and the ordering of the individuals, orders all $2n$ genes. Next we label the first gene as $s_{11} = 1$. Suppose that the first r genes have been labelled and comprise j distinct (i.e., not IBD) genes. If the $(r + 1)$st gene is IBD to any previous gene we give it the same label as that gene, otherwise label it $j + 1$. We continue until all $2n$ genes are labelled. Following Thompson (1974), we call s a *gene-identity state*. The number d of labels is the number of genetically distinct genes.

To simplify discussion we focus on noninbred individuals, although the results apply more generally. Let $\mathcal{S}_n$ denote the set of all possible gene-identity states among any n noninbred individuals. For example, Table 2 shows the seven gene-identity states in $\mathcal{S}_2$. We shall identify any two states s and s' in $\mathcal{S}_n$ that differ only in the order of paternal and maternal genes for one or more individuals. Thus in Table 2 we identify $s = (1, 2, 1, 2)$ and $s' = (1, 2, 2, 1)$, because s' is obtained from s by interchanging the maternal and paternal genes of either the first or the second individual, and relabelling according to the scheme described above. Similarly, states $s_1 = (1, 2, 1, 3)$, $s_2 = (1, 2, 3, 1)$, $s_3 = (1, 2, 3, 2)$, and $s_4 = (1, 2, 2, 3)$ all are equivalent, because all can be obtained from s_1 by a suitable interchange of maternal and paternal genes of one or more individuals and relabelling. In general, two states s and s' having the same number d of gene labels are said to be equivalent if for some permutation σ of the set $\{1, \ldots, d\}$ of labels the sets $\{s'_{i1}, s'_{i2}\}$ and $\{\sigma(s_{i1}), \sigma(s_{i2})\}$ are equal, $i = 1, \ldots, n$. This relation is an equivalence relation that partitions $\mathcal{S}_n$ into equivalence classes (Thompson, 1974). An IBD configuration ϕ is an equivalence class of gene-identity states. We write $\phi = [s_{11}s_{12} \cdots s_{n1}s_{n2}]$, where $(s_{11}, s_{12}, \ldots, s_{n1}, s_{n2})$ is any representative of ϕ. Table 2 shows the three IBD configurations possible among $n = 2$ noninbred individuals. Table 3 shows the 16 IBD configurations possible among $n = 3$ noninbred individuals.

Without loss of generality, we assume that each pedigree member has either both or none of his parents in the pedigree. Those individuals with no parents in the pedigree, called *founders*, are assumed to be unrelated and noninbred. Following Lange and Elston (1975) and Cannings et al.

Table 2
The seven gene-identity states and three IBD *configurations for n* = 2 *noninbred individuals*

		$P(\phi \mid \mathcal{R})$	
Gene identity state	IBD configuration ϕ	Full sibs	First cousins
(1, 2, 1, 2) (1, 2, 2, 1)	[1212]	$\frac{1}{4}$	0
(1, 2, 1, 3) (1, 2, 3, 1) (1, 2, 3, 2) (1, 2, 2, 3)	[1213]	$\frac{1}{2}$	$\frac{1}{4}$
(1, 2, 3, 4)	[1234]	$\frac{1}{4}$	$\frac{3}{4}$

Table 3
The 16 IBD *configurations for* $n = 3$ *noninbred individuals*

	IBD configuration ϕ	$P(\phi \vert \mathcal{P})^a$
1.	[12 12 12]	0
2.	[12 13 12]	0
3.	[12 13 13]	0
4.	[12 12 13]	$\frac{1}{8}$
5.	[12 13 23]	0
6.	[12 12 34]	$\frac{1}{8}$
7.	[12 13 34]	$\frac{1}{8}$
8.	[12 34 12]	0
9.	[12 34 13]	0
10.	[12 34 34]	0
11.	[12 13 14]	$\frac{1}{8}$
12.	[12 13 24]	0
13.	[12 13 45]	$\frac{1}{4}$
14.	[12 34 15]	0
15.	[12 34 35]	$\frac{1}{8}$
16.	[12 34 56]	$\frac{1}{8}$

[a] For the ordered set $A = \{$a man, his brother, his brother's grandson$\}$.

(1978), we represent the pedigree by a graph. An individual is represented by a node ∘, a marriage by a node •, and arrows (i.e., directed arcs) connect an individual to his marriage(s) and a marriage to its offspring (Figure 1a). We begin by assuming that the pedigree has no loops, i.e., there is no node that is connected by arcs to itself. Later we show in an example how the algorithm can be extended to arbitrary pedigrees via methods analogous to those of Lange and Elston (1975) and Cannings et al. (1978). A *nuclear family* consists of two parents, their marriage node, their offspring, and the arcs connecting them. A nuclear family is *peripheral* if only one member (parent or offspring) of the family is also a member (parent or offspring) of another nuclear family. We call this member the *pivot* of the peripheral nuclear family.

We now describe an algorithm for obtaining all IBD configurations of a set $\mathcal{A}$ of pedigree members. Before applying the algorithm, we delete from the pedigree all individuals except members of $\mathcal{A}$ and pivots on the path connecting pairs of members of $\mathcal{A}$. The first ("peeling") calculation successively peels nuclear families from the pedigree while storing their configurations. Specifically, we choose an arbitrary peripheral nuclear family to be peeled and attach to its pivot P the set of all possible gene-identity states for the family members who belong to $\mathcal{A}$. If any family member has served as a pivot previously, we expand each current state to include each of the states attached to that person. We next "peel" the family from the pedigree by removing all family members except P. Then we choose another peripheral family for peeling and repeat the process. After all nuclear families have been peeled to a single nuclear family, the set of expanded states is trimmed by deleting pivots who do not belong to $\mathcal{A}$. In the second ("pooling") calculation the remaining states are pooled into IBD configurations, and the relative frequencies of states belonging to a given configuration give the probability of that configuration. The following paragraphs provide details.

2.1 *Peeling*

Choose a peripheral nuclear family. (Since the pedigree has no loops it has at least one such family.) Each of the family's k offspring has one of four possible gene assignments (one of two from each parent). Thus for each parental state there are 4^k possible states for the offspring. If any two states differ only by a transposition of gene labels for a founder, delete one of them. List all the remaining states. Next, if anyone has served as pivot, then in each current state replace the two labels for this individual with all those in the states attached to him and relabel the (expanded) state. Attach to the

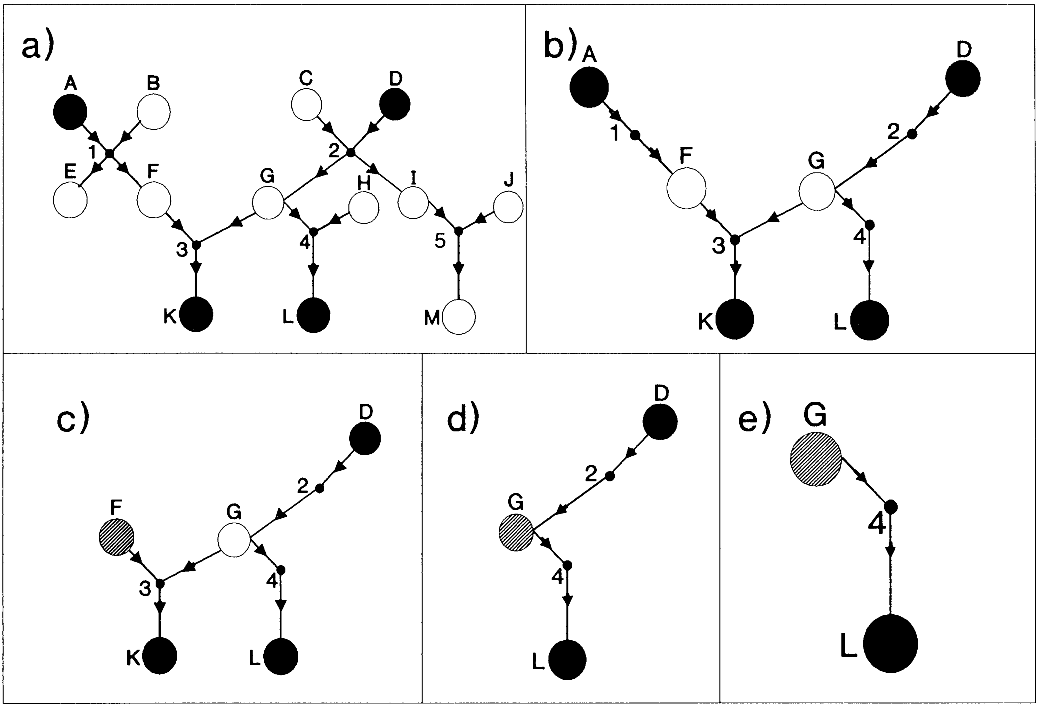

Figure 1. (**a**) A pedigree consisting of five nuclear families, labelled 1–5, with set $\mathcal{A}$ = {A, D, K, L} of affected persons. (**b**) The pedigree after deleting all individuals who neither belong to $\mathcal{A}$ nor connect members of $\mathcal{A}$. (**c**)–(**e**) The pedigree after peeling families 1 (pivot F), 3 (pivot G), and 2 (pivot G), respectively.

current pivot the new expanded states, and then "peel" the family from the pedigree, except for its pivot. Repeat this process until the pedigree consists of a single nuclear family. Then delete from each state the labels for all pivots who do not belong to $\mathcal{A}$ and relabel to obtain the final states.

We illustrate the peeling by applying it to the pedigree in Figure 1a. This pedigree consists of five nuclear families represented by marriage nodes 1, 2, 3, 4, 5. We have assumed that all 12 genes associated with founders A, B, C, D, H, and J are genetically distinct. The filled circles denote members of the set $\mathcal{A}$ = {A, D, K, L} whose IBD configuration is needed. We begin by deleting all but affected individuals and the pivots connecting them (Figure 1b). Starting arbitrarily with the peripheral family 1 having pivot F, we list the two states s_1 = (1, 2, 1, 3) and s_2 = (1, 2, 2, 3) for this family. These two states differ only by a transposition of genes for a founder (individual A) and

Table 4
Peeling peripheral nuclear families from pedigree of Figure 1

Family 1	Family 3		Family 2		Family 4		Final states
List states A F	List states F G K	Add states of family 1 attached to F A F G K	List states D G	Add states of family 3 attached to G A F D G K	List states G L	Add states of family 2 attached to G A F D G L K	Delete unaffected members and relabel A D L K
1213	123413	12134514	1213	1213464514	1213	121346454714	12343513
	123414	12134515		1213464515	1223	121346454715	12343516
	123423	12134534		1213464534		121346454734	12343563
	123424	12134535		1213464535		121346454735	12343567
						121346455714	12345613
						121346455715	12345615
						121346455734	12345673
						121346455735	12345675

thus one of them is redundant. We discard s_2 and retain s_1 in column 1 of Table 4. We now peel family 1 down to F, attaching the state s_1 to F. We arbitrarily choose family 3 as the next family to be peeled; it has the four states shown in column 2 of Table 4. Note that in the peeled pedigree shown in Figure 1c, family 3 is peripheral with pivot G. First we append to each of its four states the state attached to F, relabel the resulting states as shown in column 3 of Table 4, and attach those states to the pivot G. We then peel family 3 to obtain Figure 1d. Working next on family 2 with pivot G, we eliminate one of its two states (since they differ only by a transposition of genes for the founder D) and combine the remaining state (column 4) with the four states in column 3 that were attached to G in the previous step. This gives the four states in column 5, which we attach to the pivot G. Similarly, we combine these four states with the two states for {G, L} of family 4 shown in column 6 to obtain the eight expanded states in column 7 for the single nuclear family 4 of Figure 1e. After deleting nonmembers of $\mathscr{A}$ and relabelling, we are left with the eight states for {A, D, L, K} shown in column 8 of the table.

2.2 *Pooling*

We identify equivalent gene-identity states and pool them to obtain the probability of their IBD configuration. Suppose that a state s labels d genetically distinct genes shared by n individuals. For example, the state $s = (1, 2, 3, 4, 3, 5, 1, 6)$ labels $d = 6$ distinct genes for $n = 4$ individuals. For the νth gene, we construct a binary number $t_\nu = \tau_{\nu 1}\tau_{\nu 2} \cdots \tau_{\nu n}$, where $\tau_{\nu i} = 1$ if gene ν is carried by individual i and $\tau_{\nu i} = 0$ otherwise. Thus for s, $t_1 = 1001 = 9$, $t_2 = 1000 = 8$, $t_3 = 0110 = 6$, $t_4 = 0100 = 4$, $t_5 = 0010 = 2$, and $t_6 = 0001 = 1$. Let $t(s) = (t_{(1)}, \ldots, t_{(d)})$, where $t_{(\nu)}$ is the νth largest of the t_ν, $\nu = 1, \ldots, d$. So for our example, $t(s) = (9, 8, 6, 4, 2, 1)$. We now show that two states s and s' are equivalent if and only if $t(s) = t(s')$. Recall that s and s' are equivalent if and only if for some permutation σ of their d gene labels, $\{s'_{i1}, s'_{i2}\} = \{\sigma(s_{i1}), \sigma(s_{i2})\}$, for $i = 1, \ldots, n$. But this means that, writing each t_ν in its binary representation as a column vector, the $n \times d$ matrices $[t_1 \cdots t_d]$ and $[t'_1 \cdots t'_d]$ differ only by a permutation σ^* of their columns, given by $t'_\nu \equiv \sigma^*(t_\nu) = t_{\sigma(\nu)}$, $\nu = 1, \ldots, d$. Such a permutation exists if and only if $t(s) = t(s')$. Consider, for example, the two states $s = (1, 2, 3, 4, 3, 5, 1, 6)$ and $s' = (1, 2, 3, 4, 3, 5, 2, 6)$ for individuals A, B, C, D. Since s' is obtained from s by interchanging the paternal and maternal genes of individual A and relabelling, these two states are equivalent. We have seen that for s, $(t_1, t_2, t_3, t_4, t_5, t_6) = (9, 8, 6, 4, 2, 1)$. The same holds for s' except that the values for t_1 and t_2 are interchanged. Thus $t(s') = t(s)$ as required. Conversely, s is not equivalent to $s'' = (1, 2, 3, 4, 5, 6, 2, 5)$, since $t(s'') = (9, 8, 4, 4, 3, 2) \neq t(s)$.

This rule and appropriate sorting routines permit rapid assortment of the final states into their IBD configurations. Since each of the final states is equally likely, the relative frequencies of states in an IBD configuration give its probability. Each of the final eight states in Table 4 for the pedigree members of Figure 1 represents a unique configuration; thus each possible IBD configuration has probability $\frac{1}{8}$.

A computer program that produces the final states, the IBD configurations, and their probabilities for pedigrees without loops is available from the authors. This program required 2 minutes on a Sun-4.2 SPARC station to produce 32,768 states and 32,768 IBD configurations for a set $\mathscr{A}$ containing 12 individuals from a pedigree of 18 individuals. In general, the program requires $8xy$ bytes of memory, where x is the final number of states and y is the number of individuals who belong to $\mathscr{A}$ or serve as pivots. The final number of states is $x = \Pi_{f=1}^{F} 4^{c_f}/2^{2-p_f+h_f}$. Here F is the number of nuclear families in the reduced pedigree after deleting all but affected individuals and the pivots connecting them, p_f and c_f are, respectively, the number of parents and children in family f who either belong to $\mathscr{A}$ or are pivots, and h_f is the number of founders in f who belong to $\mathscr{A}$. In particular, the pedigree with 32,768 states required 3,145,728 bytes of memory. So although some efficiencies are possible in the algorithm's implementation, memory and storage needs limit its utility for complex pedigrees.

2.3 *Pedigrees with Loops*

Finally, we illustrate how the approach can be extended to pedigrees with loops by considering the one shown on the left in Figure 2. This pedigree consists of two nuclear families represented by the marriage nodes 1 and 2. It contains the loop consisting of individuals B and C, marriage nodes 1 and 2, and the arrows between them. We have assumed that all four genes of the founders A and B are distinct. We first break the loop by introducing a "twin" for B (Lange and Elston, 1975), i.e., an individual B′ with the same genes and pedigree position as B, as shown on the right in Figure 2. (The loop also could be broken by creating a twin for C.) Then we use the peeling part of the algorithm to enumerate all gene-identity states for the affected pedigree members. Now, however, twins are retained at each step, along with members of $\mathscr{A}$ and pivots, and whenever states are attached to one

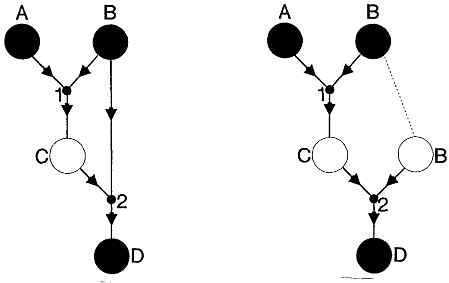

Figure 2. A simple pedigree containing a loop (left). The loop is broken by creating a "twin" B′ for B (right).

twin they also are attached to the other. Moreover in editing states, we identify the labels of twins, in this case B and B′. In the final step we delete C and B′ and relabel.

3. Applications

3.1 *Evaluating Phenotype Probabilities*

We first use the IBD configuration probabilities to evaluate phenotype probabilities. A phenotype for an individual might be a continuous or categorical trait, or an unordered pair of alleles at the given locus—the genotype. This problem has been addressed for pairs of relatives by several authors (e.g., Denniston, 1974; Karigl, 1982; Amos, Dawson, and Elston, 1990). We represent the joint phenotype for n individuals as $Y = (Y_1, \dots, Y_n)$, where Y_i denotes the phenotype of the ith individual, $i = 1, \dots, n$. We assume that Y depends on the individuals' relationship only through the IBD configuration ϕ of genes at the locus. Y also may depend on a matrix $z = (z_1, \dots, z_n)$, where z_i is a column vector of covariates for the ith individual, $i = 1, \dots, n$. Thus we write the probability of Y, conditional on the individuals' relationship $\mathfrak{R}$ and covariates z, as

$$P(Y|\mathfrak{R}, z) = \sum_{\phi} P(\phi|\mathfrak{R})P(Y|\mathfrak{R}, \phi, z) = \sum_{\phi} P(\phi|\mathfrak{R})P(Y|\phi, z). \tag{1}$$

Having determined the $P(\phi|\mathfrak{R})$ by the methods of the previous section, we must specify the probabilities $P(Y|\phi, z)$. To do so, let $g = (\{a_{11}, a_{12}\}, \dots, \{a_{n1}, a_{n2}\})$, where $\{a_{i1}, a_{i2}\}$ denotes the ith individual's genotype, i.e., the set of his maternal and paternal alleles at the locus. We assume that $P(Y|\phi, z)$ depends on ϕ through the genotypes g associated with ϕ. We also assume that allele occurrence in the n individuals does not depend on their covariates z. Then

$$P(Y|\phi, z) = \sum_{g} P(g|\phi)P(Y|g, z), \tag{2}$$

where g runs through the genotypes compatible with ϕ. To evaluate the $P(g|\phi)$, suppose ϕ contains d distinct genes, labelled $1, 2, \dots, d$ in the order specified by an arbitrarily chosen representative s. Each sequence $\alpha = (\alpha_1, \dots, \alpha_d)$ of alleles assigns a genotype $g = \alpha(\phi)$ to ϕ by the rule $\alpha(\phi) = (\{\alpha_{s_{11}}, \alpha_{s_{12}}\}, \dots, \{\alpha_{s_{n1}}, \alpha_{s_{n2}}\})$. For instance, with $n = 2$ individuals and $\phi = [1213]$, the sequence $\alpha = (a_1, a_2, a_2)$ assigns to ϕ the genotype $\alpha(\phi) = (\{a_1, a_2\}, \{a_1, a_2\})$, while $\alpha = (a_1, a_2, a_1)$ gives $\alpha(\phi) = (\{a_1, a_2\}, \{a_1, a_1\})$. Moreover any genotype g compatible with ϕ can be generated by some allele sequence α. Thus letting $P(a_i)$ denote the population frequency of allele a_i, $i = 1, \dots, m$, and assuming independence of allele occurrence at distinct genes, we rewrite (2) as

$$P(Y|\phi, z) = \sum_{\alpha} P(\alpha_1) \cdots P(\alpha_d)P[Y|\alpha(\phi), z], \tag{3}$$

where the summation is taken over all allele sequences α of length d. Substitution of (3) into (1) yields the joint phenotype distribution

$$P(Y|\mathfrak{R}, z) = \sum_{\phi} P(\phi|\mathfrak{R}) \sum_{\alpha} P(\alpha_1) \cdots P(\alpha_d)P[Y|\alpha(\phi), z]. \tag{4}$$

Existing algorithms for computing phenotype probabilities in pedigrees use not (4) but rather

$$P(Y|\mathfrak{R}, z) = \sum_{g} P(g|\mathfrak{R})P(Y|g, z), \tag{5}$$

obtained by substituting (2) into (1) and interchanging the order of summation (e.g., Elston and Stewart, 1971, pp. 524–528). This approach may be optimal for computing the probability of a single Y. However if several phenotype probabilities are needed simultaneously (as when calculating the test statistic in Example 1 below), it may be more efficient to compute them using (4) instead of (5). Further work is needed to determine optimal computing strategies for a given pedigree and application.

Two examples illustrate the use of (4).

Example 1. Linkage analysis is used to test the hypothesis that a "marker" gene of known location is distant (e.g., on a different chromosome) from a gene suspected to exist and to govern a certain disease. To illustrate, suppose the disease occurs in n relatives. The observed data are two alleles at the marker locus for each relative, i.e., the marker genotypes $g = (g_1, \ldots, g_n)$. Thus $Y = g$, and $P(Y|g, z) \equiv P(Y|g) = 1$ if $Y = g$, and $P(Y|g, z) = 0$ otherwise. Then (4) becomes

$$P(Y|\mathfrak{R}, z) \equiv P(Y|\mathfrak{R}) = \sum_{\phi} P(\phi|\mathfrak{R}) \sum_{\alpha;\alpha(\phi) = Y} P(\alpha_1) \cdots P(\alpha_d). \tag{6}$$

If the null hypothesis is false and the marker is close to the disease gene, then the marker alleles Y tend to show greater similarity across affected pedigree members than expected on the basis of the individuals' relationship. Several authors (e.g., Green and Woodrow, 1977; Weeks and Lange, 1988; Whittemore and Halpern, 1994) have proposed scoring functions (based on "identity by states") for the possible genotypes Y, with high scores given to those showing extensive allele similarity. The observed score is compared to its distribution under the null hypothesis of no linkage between marker and disease gene. This distribution is determined by (6). See Whittemore and Halpern (1994) for application of IBD probabilities to linkage analysis.

Example 2. Suppose we want the probability at birth that individual 1 develops a disease such as prostate cancer, given that the disease has occurred in two of his relatives, say his maternal grandfather 2 and maternal great-uncle 3. (We regard the genotypes of his unaffected male relatives as uninformative for this disease, which seldom occurs until after age 60 years.) Let Y_i be an indicator for disease occurrence in individual i, with $Y = (Y_1, Y_2, Y_3)$. We wish to determine

$$P(E|E_+, \mathfrak{R}) = P(E|\mathfrak{R})/P(E_+|\mathfrak{R}),$$

where E and E_+ are the events $Y = (1, 1, 1)$ and $(Y_2, Y_3) = (1, 1)$, respectively, and $\mathfrak{R}$ denotes the relationship of the three individuals. We first assume that the probability that $Y_i = 1$ depends only on g_i, and that, given the unobserved genotypes $g = (g_1, g_2, g_3)$, Y_1, Y_2, and Y_3 are mutually independent. Then (4) becomes

$$P(Y|\mathfrak{R}, z) \equiv P(Y|\mathfrak{R}) = \sum_{\phi} P(\phi|\mathfrak{R}) \sum_{\alpha} P(\alpha_1) \cdots P(\alpha_d) \prod_{i=1}^{n} P(Y_i|\alpha^{(i)}(\phi)), \tag{7}$$

where $\alpha^{(i)}(\phi) = \{\alpha_{s_{i1}}, \alpha_{s_{i2}}\}$ is the genotype assigned to the ith individual by the allele sequence α. Suppose that genetic susceptibility to the disease is governed by a single dominant allele a_1 having population frequency p, and let a_2 represent the collection of normal alleles with frequency $q = 1 - p$. Table 2 gives the probabilities $P(\phi|\mathfrak{R})$ for the brothers 2 and 3. Using these in (7) yields

$$\begin{aligned} P(E_+|\mathfrak{R}) &= \tfrac{1}{4} P(E_+|[1212]) + \tfrac{1}{2} P(E_+|[1213]) + \tfrac{1}{4} P(E_+|[1234]) \\ &= \tfrac{1}{4} [q^2\pi^2 + (1 - q^2)(r\pi)^2] + \tfrac{1}{2} [p(r\pi)^2 + q(pr\pi + q\pi)^2] + \tfrac{1}{4} [q^2\pi + (1 - q^2)r\pi]^2, \end{aligned}$$

where $r\pi$ and π are the lifetime disease risks in carriers and noncarriers of a_1, respectively. Similarly, using in (7) the $P(\phi|\mathfrak{R})$ obtained from Table 3 gives

$$P(E|\mathfrak{R}) = \tfrac{1}{8} [P(E|\phi_4) + P(E|\phi_6) + P(E|\phi_{11}) + P(E|\phi_{12}) + 2P(E|\phi_{13}) + P(E|\phi_{14}) + P(E|\phi_{16})],$$

where, for example,

$$P(E|[121213]) = \pi^3[q^3 + pq^2(r + r^2 + r^3) + 3p^2qr^3 + p^3r^3].$$

If the disease-susceptibility allele is in Hardy–Weinberg equilibrium in the population, then the proportion of carriers is $p^2 + 2pq$. Since the lifetime risk of prostate cancer in the general U.S. white male population is about .088, the assumption of Hardy–Weinberg equilibrium gives the risk π in noncarriers, in terms of p and r, as the solution to $(p^2 + 2pq)r\pi + q^2\pi = .088$, or $\pi = .088[(p^2 + 2pq)r + q^2]^{-1}$. This and the preceding equations imply that $P(E|E_+, \mathcal{R}) = .092$ if $p = .001$, $r = 10$, and $P(E|E_+, \mathcal{R}) = .106$ if $p = .01$, $r = 3$. In practice, the gene frequency p and relative risk r are estimated rather than known. The variance of $P(E|E_+, \mathcal{R})$ can be estimated from variance estimates for p and r by the delta method. Covariates $z = (z_1, z_2, z_3)$ that affect prostate cancer risk can be included by rewriting (7) as

$$P(Y|\mathcal{R}, z) = \sum_{\phi} P(\phi|\mathcal{R}) \sum_{\alpha} P(\alpha_1) \cdots P(\alpha_d) \prod_{i=1}^{n} P(Y_i|\alpha^{(i)}(\phi), z_i),$$

and specifying a model for $P(Y_i|g_i, z_i)$. For example, a linear logistic model gives

$$\text{logit}\,[P(Y_i|g_i, z_i)] = \beta_0 + \beta_1 s(g_i) + \beta_2 z_i, \quad i = 1, \ldots, n,$$

where $\text{logit}(P) = \log[P/(1 - P)]$, and $s(g_i) = 1$ if the genotype g_i involves the allele a_1 and $s(g_i) = 0$ otherwise (see also Bonney, 1986). The previous model without covariates is a special case, with $\beta_2 = 0$ and $\beta_1 = \log[r(1 - \pi)/(1 - r\pi)]$.

3.2 *Evaluating Generalized Kinship Coefficients*

Karigl (1981, 1982) and Weeks and Lange (1988) introduced generalized kinship coefficients to specify genealogical relationships in applications. A generalized kinship coefficient is the probability $P(\mathcal{P}|\mathcal{R})$ of a partition $\mathcal{P}$ of the set of genes obtained by selecting one at random from each of n individuals having relationship $\mathcal{R}$, and defining two genes in the set to be equivalent if they are IBD. Although Karigl's and Weeks and Lange's generalized kinship coefficients both specialize to the kinship coefficients of Jacquard (1972) when $n = 2$, in general they are different. Both groups of investigators have developed recursions to compute their coefficients. Here we show how to compute them as linear functions

$$P(\mathcal{P}|\mathcal{R}) = \sum_{\phi} P(\mathcal{P}|\phi)P(\phi|\mathcal{R}) \tag{8}$$

of the IBD configuration probabilities, where $\sum_\phi$ denotes summation over configurations ϕ consistent with the relationship $\mathcal{R}$.

Consider, for example, the generalized kinship coefficients of Weeks and Lange for three noninbred individuals, labelled u, v, w. There are five possible partitions of the set $\{g_u, g_v, g_w\}$, where, say, g_u denotes one of the two genes randomly selected from individual u. The five partitions $\mathcal{P}$ are $[(g_u g_v g_w)]$, $[(g_u g_v)(g_w)]$, $[(g_u g_w)(g_v)]$, $[(g_u)(g_v g_w)]$, and $[(g_u)(g_v)(g_w)]$. The corresponding five kinship coefficients are computed from (8) with summation over the 16 configurations ϕ listed in Table 3, and with $P(\mathcal{P}|\phi)$ obtained by considering the $2^3 = 8$ equally likely ways of choosing a gene from each of the three individuals.

As an example, for $\mathcal{P} = [(g_u g_v g_w)]$, we have $P(\mathcal{P}|[121212]) = \frac{1}{4}$, $P(\mathcal{P}|\phi) = \frac{1}{8}$ when $\phi = [121312]$ or $[121313]$ or $[121213]$, and $P(\mathcal{P}|\phi) = 0$ otherwise. Thus when u and w are, respectively, the sibling and grandchild of v, inserting these values for $P(\mathcal{P}|\phi)$ and the values $P(\phi|\mathcal{R})$ from column 3 of Table 3 gives $P([(g_u g_v g_w)]|\mathcal{R}) = \frac{1}{64}$. Karigl's coefficients of kinship depend linearly on those of Weeks and Lange; for example, Karigl's coefficient $\Phi_{uv} = P\{[(g_u g_v g_w)] \cup [(g_u g_v)(g_w)]\}$. Thus Karigl's coefficients also are linear functions of the $P(\phi|\mathcal{R})$.

Conversely, rather than calculate directly the IBD configuration probabilities, we could in principle use the recursions of Karigl or Weeks and Lange to compute their kinship coefficients, and then solve the resulting system of linear equations (8) for the unknown $P(\phi|\mathcal{R})$. However the large number of equations and unknowns makes this method impractical for $n > 3$.

Acknowledgements

This research was supported by NIH Grants CA 47448 and GM 21215. We thank the referees for comments that greatly improved an earlier version of this paper.

Résumé

Deux gênes en un locus donné sont identiques par descendance (IBD) si les deux proviennent d'un ancêtre commun. Nous présentons un algorithme de calcul des probabilités de toutes les relations

IBD parmi les gênes des membres d'un pedigree. Nous montrons comment utiliser ces probabilités pour calculer la probabilité d'une combinaison de génotypes ou phénotypes pour les membres d'un pedigree. Nous illustrons par des exemples, les applications à l'analyse de linkage et à la consultation génétique. L'algorithme peut également être utilisé pour calculer les coefficients de parenté généralisés proposés par d'autres auteurs.

References

Amos, C. I., Dawson, D. V., and Elston, R. C. (1990). The probability determination of identity-by-descent sharing for pairs of relatives from pedigrees. *American Journal of Human Genetics* **47,** 842–853.

Boehnke, M. (1991). Estimating allele frequency. *American Journal of Human Genetics* **48,** 22–25.

Bonney, G. E. (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics* **42,** 611–625.

Cannings, C., Thompson, E. A., and Skolnick, M. H. (1978). Probability functions on complex pedigrees. *Advances in Applied Probability* **10,** 26–61.

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

Denniston, C. (1974). An extension of the probability approach to genetic relationships, one locus. *Theoretical Population Biology* **6,** 58–75.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21,** 523–542.

Ethier, S. N. and Hodge, S. E. (1985). Identity-by-descent analysis of sibship configurations. *American Journal of Medical Genetics* **22,** 263–272.

Green, J. R. and Woodrow, J. C. (1977). Sibling method for detecting HLA-linked genes in disease. *Tissue Antigens* **9,** 31–35.

Jacquard, A. (1972). Genetics information given by a relative. *Biometrics* **28,** 1101–1114.

Karigl, G. (1981). A recursive algorithm for the calculation of identity coefficients. *Annals of Human Genetics* **45,** 299–305.

Karigl, G. (1982). A mathematical approach to multiple genetic relationships. *Theoretical Population Biology* **21,** 379–393.

Lange, K. and Elston, R. C. (1975). Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Human Heredity* **25,** 95–105.

Li, C. C. and Sacks, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10,** 347–360.

Malecot, G. (1948). Les mathematiques de l'heredité. Paris: Masson et Cie.

Thompson, E. A. (1974). Gene identities and multiple relationships. *Biometrics* **30,** 667–680.

Weeks, D. E. and Lange, K. (1988). The affected pedigree member method of linkage analysis. *American Journal of Human Genetics* **42,** 315–326.

Whittemore, A. S. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* **50,** 118–127.

Received July 1991; *revised June and December* 1992; *accepted December* 1992.