Two-locus and three-locus gene identity by descent in pedigrees

E. A. Thompson

Department of Statistics GN-22,
University of Washington,
Seattle, WA 98195.

First draft: January 1988
Second draft: March 1988

Summary

Although there have been several mathematical formulations of multilocus segregation, multilocus gene identity by descent in pedigrees has been little considered. Here we present a computationally feasible algorithm for the computation of two-locus kinship for individuals between whom there may be multiple complex relationships, and use it to investigate patterns of two-locus gene identity by descent for some standard relationships. We also present an explicit formula, which is used to discuss the determinants of two-locus identity and the relationship to 3-locus identity by descent. With the current increasing density of information on individuals genomes available from DNA polymorphisms, gene identity at linked loci has practical importance. Procedures for the estimation of relationships between individuals on the basis of genetic data will have increased flexibility to discriminate wider classes of genealogical relationship where information on multiple linked loci can be employed. Gene identity by decent at linked loci is also a key aspect of mapping rare recessive diseases from data on inbred individuals,

Introduction

There have been several mathematical formulations of multilocus segregation, originating with the paper of Geiringer (1944). Karlin and Lieberman (1979) consider the recombination process and resulting gametic distributions. Holgate (1981) formalises the gametic outputs in the framework of genetic algebras. Christiansen (1987) considers these gametic outputs with a view to analysis of population linkage disequilibria. However, multilocus gene identity by descent in pedigrees has been little considered, although Weir and Cockerham (1969) and Cockerham and Weir (1977) gave some analysis of simple systems. One reason for this has been the computational complexity of the problem; Denniston (1975) enumerated the many possible states of gene identity between two individuals at two loci. Another reason may have been the absence of data that could require consideration of linked loci. However, with the advent of DNA polymorphism data and the increasing density of information on individual genomes, this is no longer the case. Both in the estimation of genealogical relationships from genetic data (Thompson and Meagher, 1987) and in the linkage analysis of recessive diseases (Lander and Botstein, 1987) data at multiple tightly linked loci may soon become the rule rather than the exception, and multi-locus kinship is a key component of any assessment of statistical information in both these areas of inference. This paper presents an algorithm for computation of two-locus (in principle, multilocus) kinship, and investigates the properties of this function of recombination between loci for some specific genealogical relationships. Secondly we consider the general form of multilocus kinship, and its pedigree determinants, giving examples of genealogical relationships with identical two-locus kinship, but distinct three-locus kinship functions.

A recursive algorithm for two-locus kinship.

The first objective is to present a computationally feasible algorithm for the computation of two-locus kinship;

$$\begin{aligned} k_2(B,C) &= Pr(gametes\ segregating\ from\ each\ of\ B\ and\ C\ carry \\ &\quad genes\ identical-by-descent\ at\ both\ of\ two\ loci,\ between\ which \\ &\quad the\ recombination\ fraction\ is\ r). \end{aligned} \tag{1}$$

Genes are identical by descent (IBD) if they are copies of the same ancestral gene received via repeated segregations from some common ancestor within the defined pedigree of B and C. The same ancestor need not provide the alleles at both of the two loci, although if linkage is very tight (r very small) identity by descent at both loci will tend to result from a single ancestral chromosome carrying both alleles For example , in figure 1, B and C could provide chromosomes IBD at locus L from A_1 and at locus J from A_2, but only if a recombination occurred in both parents M_1 and M_2 of B and C. For tight linkage the chance that the IBD chromosome is a copy of a chromosome in either A_1 or A_2 is much greater.

Consider first the two well known methods of computation of kinship coefficients at a single locus

$$k_1(B,C) = Pr(homologous\ genes\ segregating\ from\ B\ and\ C\ are\ IBD). \tag{2}$$

This kinship coefficient is acccepted as the best overall measure of the closeness of a genealogical relationship between two individuals. Wright (1922) gave the classic formula

$$k_1(B,C) = \sum_{p(A)} (\tfrac{1}{2})^{\,m(p(A))} (\tfrac{1}{2}(1+f(A)). \tag{3}$$

Here summation is over all common ancestors A, and all paths $p(A)$ from B up via child-parent links to A and down via parent-child links to C not passing twice through any individual. The power $m(p(A))$ is the number of links on the path, and $f(A)$ is the inbreeding coefficient of the common ancestor A, the factor $\frac{1}{2}(1+f(A))$ being the probability that A will pass on IBD genes to two offspring. For future reference, it is convenient to think of $m(p(A))$ as being $m(p(A))-2+2$, the two links from common ancestor to two (necessarily distinct) offspring being omitted from the count (but contributing its separate factor $\frac{1}{2}(1+f(A))$, but the two segregations from the final invididuals B and C each contributing a factor. Again, in figure 1, there are two ancestors, A_1 and A_2, each providing one path, each of length 5: the kinship $k_1(B,C)$ is 1/32. Equation (3) was much used in early "loop-tracing" methods for the computation of inbreeding and kinship (for example, Stevens, 1975), but such methods become infeasible on large and complex pedigrees. With the advent of recursive programming languages the folowing equations have become the standard computational approach:

$$k_1(B,C) = k_1(C,B) \tag{4a}$$

$$k_1(B,B) = \tfrac{1}{2}(1+f(B)) \tag{4b}$$

$$f(B) = k_1(M,F) \tag{4c}$$

where M and F are the parents of B, and

$$k_1(B,C) = \tfrac{1}{2}(k_1(M,C)+k_1(F,C)) \tag{4d}$$

provided B is not C, nor ancestor of C, and

$$f(B) = k_1(B,C) = 0 \tag{4e}$$

for a founder individual B, not an ancestor of C. These equations provide a complete set which can be applied recursively up a pedigree until founders are encountered. These equations can be intuitively justified. Kinship is symmetric (equation (4a)). In the case of two genes segregating from B (equation (4b)) these are the same gene (and hence immediately IBD) with probability $\frac{1}{2}$ and are the two different genes in B also with probability $\frac{1}{2}$. These two genes are then the maternal and paternal genes of B, or a genes from M and a gene from F, and hence IBD with probability $k_1(M,F)$, also, by definition $f(B)$, the inbreeding coefficient of B (equation (4c)). For equation (4d), the gene segregating from B is with probabilty $\frac{1}{2}$ the gene recived from his mother M, and with probability $\frac{1}{2}$ the one received from his father, F.

Karigl (1981) extended equation (4) to the case of simultaneous gene identity between homologous genes segregating from a set of individuals, and Thompson (1983a) introduced "recursive descent probabilities" that enable one to compute the probability that a set of genes chosen one each from a set of individuals all descend from a hypothesised ancestral set. However, these extensions considered only homologous genes-- genes at a single locus. We now extend the same equation (4) to gene identity at two loci.

Define now the probability

$$k_2(J(A^{(i)}, B^{(j)}); L(C^{(k)}, D^{(l)}))$$

that two genes at locus J segregating from individuals A and B, to their i^{th} and j^{th} offspring respectively, are identical by descent and two genes at locus L segregating from individuals C and D, to their k^{th} and l^{th} offspring respectively, are also identical by descent. Note that, in addition to labelling individuals we must now also index the different segregations from each offspring in the current pedigree, in order that when we expand recursively we can identify those genes corresponding to a single gamete (or offspring chromosome) from those transmitted to different offspring. Then the two-locus kinship between two individuals G and H, or the two-locus inbreeding coefficient of their offspring, is

$$k_2(J(G^{(1)}, H^{(1)}); L(G^{(1)}, H^{(1)})),$$

the probability that the offspring carries IBD genes both at J and at L. The explicit notation of J and L in the probability k_2 is unecessary, but we retain it for greater clarity. Note first that, by symmetry between the individuals in a pair and between the two loci,

$$k_2(J(A, B,); L(C, D)) = k_2(J(B, A); L(C, D)) = k_2(J(C, D); L(A, B)).$$

where now we assume the segregation (or gamete) indicator is incorporated into the label for each individual. We may thus rearrange the arguments at any stage so that a given one of the up-to-four individuals is the first argument for locus J, and, if an argument for both loci, is also the first argument for locus L. This will enable us to summarise our recursions in just five equations:

Assume throughout that A is not B, C or D, nor an ancestor of any of them, and that M and F are the parents of A. The relationships or identities between individuals B, C and D are irrelevant, and we again assume that each of these labels incorporates a segregation indicator. For clarity, we make the segregation indicators for the individual A explicit, and use an indicator superscript (A) on the parent individuals M and F to indicate gametes to A. Then:

Where only one locus involves A,

$$k_2(J(A^{(1)}, B); L(C, D)) = \tfrac{1}{2}\,[k_2(J(M^{(A)}, B); L(C, D)) + k_2(J(F^{(A)}, B); L(C, D))] \tag{7}$$

$$k_2(J(A^{(1)}, A^{(2)}); L(C, D)) = \tfrac{1}{2}\,[k_1(C, D) + k_2(J(M^{(A)}, F^{(A)}); L(C, D))]: \tag{8}$$

While for a gamete from A,

$$\begin{aligned} k_2(J(A^{(1)}, B); L(A^{(1)}, C)) = \tfrac{1}{2}\,[(1-r)[k_2(J(M^{(A)}, B); L(M^{(A)}, C) + k_2(J((F^{(A)}, B); L(F^{(A)}, C))] \\ +r[k_2(J(M^{(A)}, B); L(F^{(A)}, C) + k_2(J(F^{(A)}, B); L(M^{(A)}, C))] \end{aligned} \tag{9a)}$$

whereas for different segregations at the two loci,

$$k_2(J(A^{(1)},B);L(A^{(2)},C)) = (1/4)[k_2(J(M^{(A)},B);L(M^{(A)},C) + k_2(J((F^{(A)},B);L(F^{(A)},C))$$

$$+ k_2(J(M^{(A)},B);L(F^{(A)},C) + k_2(J(F^{(A)},B);L(M^{(A)},C))] \qquad (9b)$$

For a gamete, plus an additional one-locus segregation from A,

$$k_2(J(A^{(1)},A^{(2)});L(A^{(1)},D)) = (1/4)[k_1(M^{(A)},D) + k_1(F^{(A)},D) + \qquad (10)$$

$$k_2(J(M^{(A)},F^{(A)});L(M^{(A)},D)) + k_2(J(M,F);L(F,D))]$$

while for two complete gametes,

$$k_2(J(A^{(1)},A^{(2)});L(A^{(1)},A^{(2)})) = 2\ r\ (1-r)\ k(M,F) + R\ [1 + k_2(J(M,F);L(M,F))] \qquad (11)$$

We have boundary values for a founder A, of, respectively 0, $\frac{1}{2}\ k(C,D)$, 0, 0, and R where $R = \frac{1}{2}\ (r^2 + (1-r)^2)$ for the five equations (7)-(11). These equations all derive from the same straightforward consideration of segregation probabilities as does the more familiar equation (4):

Equation (7) : A gene chosen from A at a single locus derives with probability $\frac{1}{2}$ from M and $\frac{1}{2}$ from F.

Equation (8) : If two genes segregate independently from A at locus J, they are copies of the same gene with probability $\frac{1}{2}$, and are copies of the two genes in A, and hence one from M and one from F, also with probability $\frac{1}{2}$. Note that our recursive computation of two-locus kinship will never involve an expansion for the same gamete at the same locus, although were we to consider a term such as $k_2(J(A^{(1)},A^{(1)});L(C,D))$ the expansion would be immediate, since the two genes indicated at locus L are necessarily IBD.

Equation (9) : For a single gamete segregating from A, the genes at loci J and L will both derive from M or both from F if there is no recombination (each possibility having probability $\frac{1}{2}\ (1-r)$), and from M at locus J and from F at locus L, or vice versa, if there is a recombination (each possibility having probability $\frac{1}{2}\ r$) -- hence equation (9a). On the other hand, for two separate gametes, we have the analogous gene origins in M and F, but independent segregation (equation (9b)). Alternatively, equation (9b) may be regarded as two separate applications of equation (7); both would, of course have to be completed before proceeding further up the pedigree, since we cannot expand from M or F while A remains in the expression.

Equation (10) : If genes segregate from A twice at locus J, once in combination with a gene at locus L, then r cancels from the possibilities. With probability $\frac{1}{2}$ the two genes at locus J are copies of the same gene (and hence already IBD), and then with probability $\frac{1}{2}$ the gene at L derives from M or F each with probability $\frac{1}{2}$. Alternatively, the two genes from A at locus J are genes from M and from F, while again the gene at L derives from either M or F, providing the second two terms of equation (10).

Equation (11) : Finally, equation (11) covers the case of two gametes from A. If one is recombinant and not the other (probability $2r(1-r)$) then at one locus the two genes are IBD and at the other locus they derive from M and F and are IBD with probability $k_1(M,F)$. If both chromosomes are recombinant, or both not, (probability $2R$), then with probability $\frac{1}{2}$ the genes at each locus are copies of the same parental gene and with probability $\frac{1}{2}$ we have genes from M and from F at both loci. Again, if different gametes from A are indicated ($k_2(L(A^{(1)}, A^{(2)}); J(A^{(1)}, A^{(3)}))$ for example), we can apply the recursions (7) or (9a) to the separate segregations. As in the case of equation (9b) all expansions of A's segregations must be completed before proceeding further up the pedigree.

As a check, it is worth noting that, when $r = 0$,

$$k_2(J(A,B); L(A,B)) = k_1(A,B).$$

and the above equations reduce to the single-locus equations (4). Also, when $r = \frac{1}{2}$.,

$$k_2(J(A,B); L(C,D)) = k_1(A,B)\, k_1(C,D)$$

and equations (7)-(11) give the products of two independent single-locus recursions. Note again that for the present problem, we are not concerned that the genes at loci J and L should derive from the same ancestral chromosome, although we could also derive very similar equations for this case. In practice, where linkage is tight, and/or there are few inbreeding loops, it will often be the case that where genes at both loci are identical by descent they will in fact all derive from the same ancestral chromosome, but this is not necessary. Note also that instead of gene identity probabilities, we could very similarly consider gene descent probabilities, specifying founder genes at locus J and at locus L from which descent is to be considered. Just as the single-locus gene-descent equations of Thompson (1983b) are generalisations of the single-locus gene-identity equations of Karigl (1981), the analogous multilocus generalisation can be derived here. However, we do not pursue this as there seems to be no immediate practical application.

Although some of the above equations involve a four-fold branching recursion, they are simple to implement and computationally feasible even on large and complex genealogies. In computing two-locus kinship $k_2(J(A,B); L(A,B))$ additional symmetry allows a further reduction of the problem, by combining identical terms among the parents of A and B (Table 1). Routines that trim and reorder the individuals in a pedigree to avoid lengthy unecessary parts of the recursion, similar to those used in recursive formulae for descent probabilities (Thompson, 1986) can also increase the size and complexity of the pedigree on which computation is feasible. In spite of the greater complexity of the equations owing to the genes being at two loci, it seems that computation is feasible on any pedigree on which any other four-fold gene identity can be computed.

Note also that the formulae extend (in principle) to more loci, $L_1, L_2, \cdots, L_s$. For example, we might wish to consider a s-locus expression

$$k_s(L_1(A, B_1); L_2(A, B_2); \cdots; L_s(A, B_s)),$$

where again the individual label A is assumed to include a segregation indicator, and thus a single s-locus gamete from A is implied. For any individual A, who is not the same as nor an ancestor of any other individual in the current function, we can express the s-locus kinship, k_s, as a weighted average of the multilocus kinships between the other individuals and the two parents M and F of A, the weights being simply the multilocus segregation probabilities -- the probabilities with which the relevant genes segregating from A originate from M and F. For the above particular case, provided A

is distinct from all the B_j (who need not be distinct) and not an ancestor of them, we would have

$$k_s(L_1(A,B_1); L_2(A,B_2); \cdots; L_s(A,B_s)) =$$

$$\sum_{\Delta=(\delta_1,\cdots,\delta_s)} P(\textit{gamete from A has parent origins } \Delta) \qquad (12)$$

$$k_s(L_1(\delta_1,B_1); L_2(\delta_2,B_2); ; \cdots; L_s(\delta_s,B_s))$$

where each δ_j, $j = 1, \ldots, s$ is M or F as the gene at locus j originates from M and F, and summation is over all vectors length s with components M or F. However, this expansion may involve 2^s distinct terms, and for $s > 2$ general implementation does not seem to be practicable.

Examples of gene identity by descent at two linked loci.

At a single locus, many distinct genealogical relationships have the same kinship coefficient. The sum over paths of powers of $\frac{1}{2}$ (equation (1)) can achieve a specified dyadic rational in many different combinations. For a single locus, distinct genealogical relationships between two individuals can provide not only the same kinship coefficient, but the same probabilities for patterns of gene identity by descent between the two unordered pairs of genes of the two individuals, and hence the same pairwise genotype and phenotype distributions. An example is the set of pairwise relationships grandparent-grandchild, uncle-niece, and half-sib. These three relationships can never be distinguished on the basis of genetic data at independent loci, for, regardless of the characteristics of a locus in terms of allele frequencies and the relationship between genotype and phenotype, all three provide identical pairwise phenotype distributions. On the other hand, it is known that these three pairwise relationships are, in principle, distinguishable on the basis of data at linked loci, since the probabilities that the pair have a gene in common at both loci are functions of the recombination fraction r that differ between the relationships (see for example, Thompson 1986).

It is therefore of interest to consider two-locus kinship between relationships with the same one-locus kinship, and in particular we consider the six relationships between non-inbred individuals which all have kinship coefficient 1/16. These are:

a) Great-grandparent -- great-grandchild,
b) Half-uncle -- Half-niece.
c) First cousins,
d) Double half-first cousins,
e) Quadruple second cousins (paired sibship exchange)
and f) Quadruple second cousins (cyclic sibship exchange),

The pedigrees of these six relationships are shown in figure 2, and the graphs of their two-locus kinship, as functions of the recombination fraction r, are shown in figure 3.

We note that the functions are all distinct and non-intersecting on $0 < r < \frac{1}{2}$. Not only does a greater multiplicity of ancestral loops lower the two-locus kinship, but the pattern of joining of these loops affects the result. Thus there is even distinction between the two types of quadruple second cousin. The fact that relationships are distinguished ensures their identifiability on the basis of data at pairs of linked loci. Of course, the difference is not large, and many pairs of highly informative loci would be required to discriminate reliably between the two types of quaduple second cousins, but with the increasing density of DNA polymorphisms mapped in plant and animal genomes such data are potentially available.

The other current practical application of two-locus kinship is in the mapping of rare recessive diseases on the basis of data on affected inbred individuals. The idea underlying that analysis is that the individuals will (likely) have IBD genes at the disease locus, so that the probabilities for the genotype at a linked marker locus will depend on the recombination fraction r, enabling r to be estimated or linkage inferred. This idea was originally considered by Smith (1953) and has been discussed in intervening years (Suarez et al., 1978), but until recently the power of the method was considered too slight for it to be practically relevant. Lander and Botstein (1987) point out that multilocus DNA variants may provide sufficient polymorphism for linkage to be established by this method. In their analysis, Lander and Botstein consider multiple linked loci in linkage equilibrium, but the key to the method is the degree of marker polymorphism which this assumption provides. They consider individuals with a specific inbreeding coefficient -- the parents being first cousins, or second cousins. However, the function of recombination that, given an affected individual, determines the genotype distribution at the linked marker locus is not merely the kinship of the parents (or inbreeding of the offspring), but the two-locus kinship coefficient. Generally, the smaller the two-locus kinship the lesser the power of the method (Walters, 1988).

To consider how far multiple ancestral loops can reduce two-locus kinship, we consider the case of a Hutterite sibship in which an offspring individual is affected by two very rare recessive disorders (Lowrey et al., 1985). The genealogy of this genetic isolate has been traced back eleven generations to a set of about 70 founders. Fifty-five of these are ancestors of the sibship in question, in which the offspring have inbreeding coefficient 0.099. The parents of the sibship are quadruple second cousins (of the paired exchange type) -- a regularity of ancestral relationship which is rare, even in this population of multiple sibship exchanges and prefered second-cousin marriages. Over 1000 ancestral paths connecting the parents of the individuals have been traced; the cousin relationships between the parents as given by the recursive "rels" routine of Thomas (1987) is given in table 2.

Figure 4 shows the two-locus kinship between the parents of this sibship, when the genealogy is trimmed at a specified number of generations before the present, from 4 (the quadruple second cousin relationship) to 10 (the maximum depth of known relationship) Of course, as more ancestral loops are included, any kinship coefficient between the parents increases, but for intermediate recombination fractions the two-locus kinship increases relatively less. Note that for trimming at 7 generations depth (which includes 95% of the total known inbreeding), the two-locus kinship is below that for a half-uncle-niece relationship over the range r=0.1 to 0.35. The lower part of this range, at least, is important for linkage analysis where DNA polymorphisms at recombination fractions in the range 0.1 to 0.2 are still considered closely linked (Botstein et al., 1980).

Relationships with identical two-locus and three-locus identity.

Given the power of linked loci to resolve relationships which are not identifiable from data at unlinked loci, we may ask what are the determinants of two-locus kinship, and what distinct genealogical relationships will remain indistinguishable. To answer this we return to the Wright formula for the kinship coefficient (equation (1)) and extend this to multiple loci. For any number of loci, $L_1, \ldots, L_s$, and ordered s-tuples of ancestral paths, from one individual up to a common ancestor and down to the other,

$$Pr(gametes\ from\ the\ two\ individuals\ are\ IBD\ at\ L_1, \cdots, L_s) \qquad (13)$$

$$= \sum_{(p_1,\cdots,p_k)} Pr(gametes\ are\ IBD\ at\ L_j\ via\ path\ p_j, j = 1, \ldots, s).$$

In particular, for two loci,

$$Pr(gametes\ from\ two\ individuals\ are\ IBD\ at\ locus\ J\ and\ at\ locus\ L)$$

$$= \sum_{(p,q)} Pr(gametes\ are\ IBD\ at\ J\ \ via\ path\ p\ and\ IBD\ at\ L\ via\ path\ q),$$

where ancestral paths p and q are defined as for the single-locus case. Considering the terms in this sum for which the original ancestor A at the head of path p is the same as that for path q, we have, the probability

$$(1/2)^{n(p)+n(q)-c(p,q)}(1-r)^{c(p,q)-j(p,q)}r^{j(p,q)}R_A(r, f, h(r)) \qquad (14)$$

where $c(p,q)$ is the number of parent-offspring links common to the two paths, $n(p)$ is the number in p, $n(q)$ the number in q, $j(p,q)$ the number of joins in the two paths, and the factor R_A being given below. In the count of links, we do not include the two segregations from the common ancestor A at the head of the path, but we do include the two segregations from the final individuals (see equation (1) and notes following). The rationale for formula (14) and the path counts is shown in figure 5. Where the paths are separate each segregation contributes, as for a single locus, the factor $\frac{1}{2}$, this being the probability that the "right" gene is passed on. Path p has $n(p) - c(p,q)$ such separate links, while path q has $n(q) - c(p,q)$. Where the paths have a segregation in common, we require the "right" pair of genes received from the parent to be passed to the offspring; thus, where the paths proceed together from grandparent to parent to child, there must be no recombination and we have a factor $\frac{1}{2}(1-r)$. On the other hand, where the paths join in the parent, to obtain the gene from each locus via the different paths, recombination is necessary, and we have the factor $\frac{1}{2}\,r$. Finally, where two chromosomes segregate independently from the same common ancestor, A, we have an additional factor which gives the probability that the two offspring of A share genes at both loci.

$$R_A(r, f, h(r)) = (1 - 2f + h(r))\ R(r, 0, 0) + f. \qquad (15)$$

where f is the inbreeding coefficient of A h(r) his 2-locus inbreeding, and $R(r, 0, 0)$ is the previous parameter $R = \frac{1}{2}(r^2 + (1-r)^2)$. If paths p and q are via different common ancestors A_p and A_q at the heads of the paths, the factor R_A is modified to be

$$(\tfrac{1}{2}(1 + f(A_p)))(\tfrac{1}{2}(1 + f(A_q)))$$

the probability that the two offspring of each of the two ancestors receive the same gene at the locus applicable to that path. This is the only change to equation (14).

Note that when r=0, we have only contributions from paths $p = q$, $j(p, q) = 0$, $c(p, q) = n(p) = n(q)$ and also then $h = f$, and $R_A = \frac{1}{2}(1 + f(A))$. The formula then reduces to the single locus kinship (equation (1)). When $r = \frac{1}{2}$, on the other hand, the double summation factorises, $h = f^2$, $R_A = (\frac{1}{2}(1 + f(A)))^2$ and we have separate summations over p and q giving the square of this single-locus formula.

We may similarly obtain the contributing factors for the s-locus case (equation (13)). Any segregation or pedigree link will be common to some subset of the paths. Where paths for a subset S of the s loci have a common link, the factor contributed to the term of the sum (13) expresses transmission of this locus subset without recombination. Where two paths, for sets of loci S_1 and S_2 join, the contribution is the multilocus segregation probability that in a chromosome from the offspring the locis S_1 derive from the maternal gamete and S_2 from the paternal. (Note that, since individuals have only two parents, only two sets of paths can join at a segregation). We return to this below for the case of three loci.

Thus, for two loci, any relationship for which the numbers of ancestral paths, the numbers $n(p)$, $n(q)$, $j(p, q)$ *and* $c(p, q)$ are the same, will have the same two-locus kinship. Conversely, equation (14) shows that indeed two-locus kinship will distinguish between standard relationships often considered to have the same degree of closeness, since these are obtained precisely by varying the number of ancestors and paths in a (single-locus) compensating manner-- the example of the six regular relationships of kinship 1/16 considered in the previous section is no special case. On the other hand, the two relationships of figure 6 have the same 2-locus kinship. However this example is extreme; the two pedigrees show the same s-locus kinship, for all s, or the same *genomic kinship.* Regardless of the set of loci within the genome that are considered, chromosomes from the individuals have identical similarities. They are, however, distinct genealogies, and in terms of the pairwise genotypes of the individuals they are distinguishable -- in the first case B is inbred and in the second, not. Since there are relationships with identical (single-locus) kinship but different two-locus kinship, a more interesting question is whether there are relationships of identical two-locus kinship, distinguished by three loci, and so on.

Extension of the above formulation in terms of joins and separations in ancestral paths show that this is indeed possible. The simplest examples found, of kinships distinct at three loci but not at two, involve the 3-path template shown in figure 7. Here, by "transferring" a segregation from one path segment to another, we achieve a second set of three paths, showing the same pairwise matrix of links and links in common (Table 3). Additionally, every pair of distinct paths joins once, and the (non-inbred) common ancestor at the head of every path (or path pair) contributes the same factor to the sum. Thus we have the same two-locus kinship, as has been validated numerically using the recursive algorithm. We may question why such a pair of pedigrees are distinguished by three loci, since there is but a single set of three paths. However, for three loci, locus order, and thence the order of joining of paths, is significant. Suppose the chromosomal order of three loci is L_1, L_2, L_3 and note that the order of joining of paths is p_1, p_2, p_3. Consider the cases under the two pedigrees where L_1 is taken via

p_1. If L_2 is taken via p_2, and L_3 via p_3, the order of joining of paths required first a recombinant between L_1 and L_2 and then one between L_2 and L_3. However, if L_2 is taken via path p_3 and L_3 via path p_2, we require first a recombinant between L_1 and L_3, and then a double recombinant when the path for L_2 (p_3) joins (Table 3). These two triples thus provide different contributions in terms of the recombination fractions. Since under the interchange between the two pedigrees, p_1 translates to p_3, in terms of path lengths and links in common, but the order of joining of the paths remains p_1, p_2, p_3, the three-locus kinship is different on the two pedigrees.

Another example is shown in figure 8 and table 4. This example is of interest in that under the interchange of links between two levels in the pedigree, no ancestral path changes in length: we do not have the interchange of paths that is a feature of the figure 7 template. Each row of the two matrices of path segments in common contains the same numbers, and each pair of paths joins once, providing for equal two-locus kinship. (Again, this has been validated by the recursive algorithm.) Again, however, under the interchange of links, the order of joining of analogous paths is changed, and three locus kinship is thus also changed. This pedigree is of interest also in that in two respects it demonstrates that pedigrees are very special graphs. First, and most important, individuals have only two parents. The length of the segment marked w is thus at least 1: only two (sets of) paths can join at a time. This makes balancing the numbers of links in common a more complex task, since the three-way symmetry between the paths is lost. Secondly sex must be assignable on a true pedigree. Note it is impossible therefore, that A mates with B, who mates with C who mates with A. Thus some, at least, of the arcs labelled i, j, k, l, m or n must be greater than 1. Our example numbers (Table 4) have been derived with these constraints in mind, although they are not the only possibilities.

Conclusion

With the increasing density of information on individual genomes, provided by data on DNA polymorphisms, gene identity by descent at linked loci becomes of practical importance. The facility to compute and analyse patterns of two-locus kinship in complex pedigrees, provides a new tool for examining the properties of statistical procedures for the analysis of both current and potential data. A very recent example is provided by Weeks and Lange (1988) who use multilocus kinships at $r = \frac{1}{2}$ to provide a test of departure from independent segregation. Of course, at $r = \frac{1}{2}$ gene identities at the different loci are independent. The present suggested methodology shows that it is unecessary to consider only the two extremes of complete linkage and free recombination -- analysis for arbitrary degrees of linkage is possible.

That two locus analysis is an essential generalisation of single locus analysis because of the phenomenom of linkage is well acknowledged. That three locus analysis carries an additional dimension is less widely appreciated in area of applications, although the critical relevance of locus order arises also in other problems of linkage analysis when we consider more than two loci jointly (for example, Thompson 1984). While this feature of the problem makes it intrinsically more difficult to "balance" sets of paths to achieve equal multilocus kinship, there seems no reason to suppose this impossible. While it is tempting to conjecture that "three-locus kinship-equivalence implies genomic kinship-equivalence", (a mathematical geneticists' analogue of the now class "period three implies chaos" of bifurcation theory), this seems unlikely. Rather, it seems that in a similar manner, but with more complex pedigrees, it should always be possible to construct relationships which have identical s-locus kinship, but are distinguished by (s+1)-locus kinship.

Acknowledgment

This research was supported in part by PHS grant 1R43-RR03768, and NSF grant BSR-8619760. I am grateful to Ken Morgan (McGill University) for permission to use the complex Hutterite pedigree as an example, and for his independent verification of numerical values of two-locus kinship on this pedigree.

References

Botstein, D., White, R.L., Skolnick, M.H. and Davis, R.W. 1980 Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32,** 314-331.

Christiansen, F.B. 1987 The deviation from linkage equilibrium with multiple loci varying in a stepping-stone cline. *J. Genet.* **66,** 45-67.

Cockerham, C.C. and Weir, B.S. 1977 Digenic descent measures for finite populations. *Genet. Res. Camb.* **30,** 121-147.

Denniston, C. 1975 Probability and genetic relationship: two loci. *Ann. Hum. Genet.* **39,** 89-104.

Geiringer, H. 1944 On the probability theory of linkage in Mendelian heredity. *Ann. Math. Stat.* **15,** 25-57.

Holgate, P. 1981 Population algebras. *J. Roy. Statist. Soc. (B):* **43,** 1-19.

Karigl, G. 1981 A recursive algorithm for the calculation of identity coefficients. *Ann. Hum. Genet.* **45,** 299-305.

Karlin, S. and Lieberman, U. 1979 A natural class of recombination processes and related measures of crossover interference. *Adv. Appl. Prob.* **11,** 479-501.

Lander, E.S., and Botstein, D. 1987 Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science,* **236,** 1567-1570.

Lowry, R.B., Snyder, F.F., Wesenberg, R.L., Machin, G.A., Applegarth, D.A., Morgan, K., Carter, R.J., Toone, J.R., Holmes, T.M., and Dewar, R.D. 1985 Morquio syndrome (MPS IVA) and hypophosphatasia in a Hutterite kindred. *Am. J. Med. Genet.* **22,** 463-475.

Smith, C.A.B. 1953 The detection of linkage in human genetics. *J. R. Stat. Soc. B* **15,** 153-192.

Stevens, A. 1975 An elementary computer algoritm for calculation of the coefficient of inbreeding. *Inf. Proc. Letters* **3,** 153-163.

Suarez, B.K., Rice, J., and Reich, T. 1978 The generalized sib pair IBD distribution: its use in the detection of linkage. *Ann. Hum. Genet.* **42,** 87-94.

Thomas, A.W. 1987 Pedpack -- Users' Manual. Technical report #99, Department of Statistics, University of Washington.

Thompson, E.A. 1983a A recursive algorithm for inferring gene origins. *Ann. Hum. Genet.* **47,** 143-152.

Thompson, E.A. 1983b Gene extinction and allelic origins in complex genealogies. *Proc. R. Soc. Lond. B* **219,** 241-251.

Thompson, E.A. 1984 Information for joint linkage analysis. *IMA J. Math. Appl. Med. Biol.* **1,** 31-50.

Thompson, E.A. 1986 *Pedigree Analysis in Human Genetics.* Baltimore and London: The Johns Hopkins University Press.

Walters, E. 1988 *Comparison of Linkage Analysis Designs based on Individuals Affected with Recessive Diseases.* MS. Thesis; Department of Biostatistics, University of Washington.

Weeks, D.E. and Lange, K. 1988 The affected-pedigree-member method of linkage analysis. *Amer. J. Hum. Genet.* **42,** 315-326.

Weir, B.S. and Cockerham, C.C. 1969 Pedigree mating with two linked loci. *Genetics* **61,** 923-940.

Wright, S. 1922 Coefficients of inbreeding and relationship. *Am. Nat.* **56,** 330-338.

Table 1: Initial expansion of two-locus kinship between two individuals A and B who are not ancestor and descendant using equation (9a). Symmetries then provide a reduction in the initial expansion from 16 terms to 10, which can provide a substantial saving in computing time. For those terms which remain of similar $L(A,B); J(A,B)$ form the same reduction can be made in successive expansions. For simplicity, the segregation indicators are omitted; for the initial expansion in a two-locus kinship there is only a single relevant segregation for each individual.

Individuals	Parents of individual A		Parents of both individuals	
Term	Term	Probability	Term	Probability
$L(A,B); J(A,B)$	$L(M_A,B); J(M_A,B)$	$\frac{1}{2}(1-r)$	$L(M_A,M_B); J(M_A,M_B)$	$(1-r)^2/4$
			$L(M_A,M_B); J(M_A,F_B)$	$r(1-r)/2$
			$L(M_A,F_B); J(M_A,F_B)$	$(1-r)^2/4$
	$L(F_A,B); J(M_A,B)$	r	$L(F_A,M_B); J(M_A,M_B)$	$r(1-r)/2$
			$L(F_A,F_B); J(M_A,M_B)$	$r^2/2$
			$L(F_A,M_B); J(M_A,F_B)$	$r^2/2$
			$L(F_A,F_B); J(M_A,F_B)$	$r(1-r)/2$
	$L(F_A,B); J(F_A,B)$	$\frac{1}{2}(1-r)$	$L(F_A,M_B); J(F_A,M_B)$	$(1-r)^2/4$
			$L(F_A,M_B); J(F_A,F_B)$	$r(1-r)/2$
			$L(F_A,F_B); J(F_A,F_B)$	$(1-r)^2/4$

Table 2: The relationships between the parents of a certain Hutterite individual. The table gives the number of full and half cousin relationships of each degree, and each level of removal. For example, the couple are sixth cousins once removed 126 full times and 24 half times. Their primary relationship is 4 full times (i.e. quadruple) second cousins. The total number of ancestral paths is twice 496 plus 42, or 1034.

full	half	level	removal
4	0	2	0
2	0	3	0
4	0	3	1
4	0	4	0
2	0	4	1
16	4	5	0
22	4	5	1
14	8	6	0
126	24	6	1
44	0	6	2
118	0	7	0
124	0	7	1
16	2	8	0
496	42		

Table 3: Path lengths and path-sections in common for the pedigree template of figure 7.

	General case			Pedigree I		
				$s = u = 1,\ x = t = y = 2$		
	p_1	p_2	p_3	p_1	p_2	p_3
p_1	$x + s + 2$	$s + 2$	2	5	3	2
p_2	$s + 2$	$s + u + y + 2$	$y + 2$	3	6	4
p_3	2	$y + 2$	$y + t + 2$	2	4	6

	General case			Pedigree II		
				$u = y = 1,\ x = s = t = 2$		
	p_1	p_2	p_3	p_1	p_2	p_3
p_1	$x + s + 2$	$s + 2$	2	6	4	2
p_2	$s + 2$	$s + u + y + 2$	$y + 2$	4	6	3
p_3	2	$y + 2$	$y + t + 2$	2	3	5

Note that, in both these pedigrees, the paths p_1 and p_2 join above the Note that, in both these pedigrees, the paths p_1 and p_2 join above the link of length s, and the pair are then joined by p_3 below this link. On the other hand, in the matrix of link lengths in common, the roles of p_1 and p_3 are interchanged between the two pedigrees.

Table 4: Paths sections and path lengths in common for the pedigree template of figure 8.

	General case path sections	$j = l - m = n = 1, i = k = 2, w = s = 1$ $x = y = 1, a = z = 2, b = c = 3$					
		p_1	p_2	p_3	p_4	p_5	p_6
p_1	$a + i + x + w + s + 2$	9	6	5	4	3	3
p_2	$a + k + y + w + s + 2$	6	9	4	5	2	2
p_3	$b + j + x + w + s + 2$	5	4	9	4	6	3
p_4	$c + l + y + w + s + 2$	4	5	4	9	3	6
p_5	$b + m + z + s + 2$	3	3	6	3	9	5
p_6	$c + n + z + s + 2$	3	3	3	6	5	9

	General case path sections	$j = l - m = n = 1, i = k = 2, w = s = 1$ $a = 1, b = c = x = y = 2, z = 3$					
		p_1	p_2	p_3	p_4	p_5	p_6
p_1	$a + i + x + w + s + 2$	9	5	6	4	3	3
p_2	$a + k + y + w + s + 2$	5	9	4	6	3	3
p_3	$b + j + x + w + s + 2$	6	4	9	4	5	3
p_4	$c + l + y + w + s + 2$	4	6	4	9	3	5
p_5	$b + m + z + s + 2$	3	3	5	3	9	6
p_6	$c + n + z + s + 2$	3	3	3	5	6	9

Figure captions

Figure 1: An example pedigree, showing the ancestral paths which contribute to kinship between the two individuals B and C.

Figure 2: Six relationships with single-locus kinship 0.625: (a) Great-grandparent and great-grandchild; (b) Half-uncle and half-niece; (c) First cousins; (d) Double half first cousins; (e) Quadruple second cousins (exchange type); (f) Quadruple second cousins (cyclic type).

Figure 3: Two-locus kinship as a function of recombination fraction, for the six relationships (a)-(f) of figure 2. The curves are in monotone decreasing order, with (a) giving the maximal values for each r-value and (f) the minimum.

Figure 4: Two-locus kinship as a function of recombination fraction, for the pedigree of a certain Hutterite individual, when the pedigree is trimmed at 4,5,6,7 and 10 generations before the present. The curves are in monotone increasing order, with the 4-generation trim giving the minimum values and 10 the maximum. The 4-generation trim is also the curve (e) of figure 3, and for comparison the curve (b) of figure 3 (for half-uncle/half-niece) is also shown (broken line).

Figure 5: The pattern of segregations on two paths, showing the events required for descent of requisite genes at two loci.

Figure 6: Two pedigrees showing pairs of individuals between whom there is equal genomic kinship.

Figure 7: A pedigree template, showing how equal two-locus kinship, but distinct three-locus kinship, may be obtained. If the values of y and s are interchanged (s, $y \geq 1$), and x, u and t remain fixed, the two-locus kinship is unchanged. An example set of values and resulting path lengths is given in Table 3.

Figure 8: A more complex pedigree, exhibiting a similar phenomenom to that of figure 6, but without change of any total path lengths. If each of a, b and c is increased by a specific amount, and each of x, y and z decreased by the same amount, other link lengths remaining unchanged, then, provided $x + w = z$, $a + w = c$ $b = c$, $x = y$, the two-locus kinship is unchanged. For specific example values and path lengths, see Table 4.

Erratum:

In Two-locus and three-locus gene identity by descent in pedigrees *I.M.A. J. Math. Appl. Med. and Biol.* (1988) **5:** 261-279.

In figure 3 (P.269), curve (a) is incorrect. The two-locus kinship curve for relationship (a) (great-grandparent/great-grandchild) should be identical to that for relationship (b) (half-uncle/half-niece). There is no error in the algorithm or in the computer program; the program gives the correct result. The curve shown was erroneously extracted from a different results file -- it shows the probability of a given great-grandparental chromosome segregating from the great-grandchild without recombination. I am grateful to Dr. Daniel Weeks for pointing out this error.