Proc. R. Soc. Lond. B **219**, 241–251 (1983)
Printed in Great Britain

Gene extinction and allelic origins in complex genealogies

By Elizabeth A. Thompson
Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics,
16 *Mill Lane, Cambridge CB*2 1*SB, U.K.*

With the increasing emphasis on data analysis in mathematical genetics, problems of parametrizing genealogical structure become of practical importance. A complete specification of the genetic effects of genealogical structure is provided by the probabilities of genetically distinct states of gene identity by descent. Although this provides a direct parametrization for the joint distribution of traits on a set of related individuals, it is an unwieldy tool in the analysis of large and complex genealogies. Probabilities of joint descent of founder genes and likely ancestries of alleles provide alternative characterizations of relationship and have direct application in practical problems. Joint extinction probabilities of founder genes can also be derived as ancestral likelihoods: evolutionarily, the most significant characteristic of a genealogical structure must be its effect on the survival and extinction of genes.

1. Population structure

Genetic variability is the basis of evolution, and much of the evolution of higher organisms, and especially of man, may have taken place within small isolated groups of individuals, within which short-term history may have had long-term consequences. An analysis of the structure of such groups is an important part of an understanding of the role of detailed genealogical history in the determination of current genetic distributions. Over the last few years the emphasis in mathematical genetics has moved from analyses of genetic models of evolutionary processes towards methods for the analysis of data, and thus towards more detailed descriptions of small-scale phenomena. In a small population or population sample, it is the genealogical structure which provides the essential link between observable characteristics of individuals and genetic models for the determination of such observations.

I shall restrict discussion of population structure to the context of a single Mendelian autosomal locus. That is, for the particular characteristic of interest, the type of an individual is determined by the types of the two genes that he carries, one of which he received from his father and the other from his mother; to each of his offspring he will pass on a randomly chosen one of these two genes. A *gene* in an individual will refer to one of these two homologous genes, and a *trait* will be an observable characteristic of individuals determined by the unordered pair of types of these two genes. Of course, evolutionary processes involve very much more than the segregation of discrete Mendelian autosomal genes: but, whatever the ramifications of DNA sequences and complex multi-locus systems, it remains the fact that much of the normal variation observed within populations is of

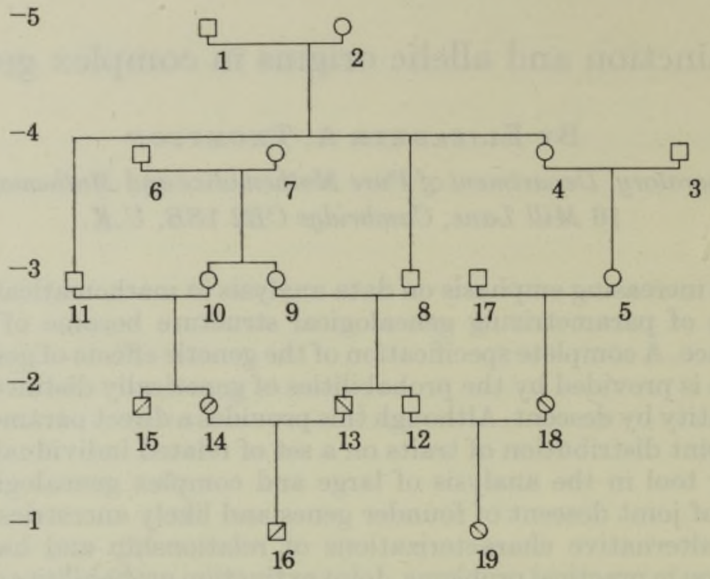

FIGURE 1. Genealogy used for the purposes of example throughout this paper. Males are denoted by squares and females by circles. Oblique strokes denote current and current carrier individuals (see table 1).

TABLE 1. SPECIFICATION OF THE EXAMPLE GENEALOGY OF FIGURE 1

individual	mother	father	sex	comment
1	0	0	1	founder
2	0	0	2	founder
3	0	0	1	founder
4	2	1	2	—
5	4	3	2	—
6	0	0	1	founder
7	0	0	2	founder
8	2	1	1	—
9	7	6	2	—
10	7	6	2	—
11	2	1	1	—
12	9	8	1	—
13	9	8	1	current
14	10	11	2	current carrier
15	10	11	1	current carrier
16	14	13	1	current carrier
17	0	0	1	founder
18	5	17	2	current
19	18	12	2	current

variants without marked selective effects, not closely linked to other markers, and segregating according to Mendel's first law. Such traits are involved in many of the open questions of data analysis.

In principle, a genealogy is a graph with some special characteristics. Everyone has precisely two parents, and a specification of the parents of all individuals past and present *is* the genealogy (see figure 1 and table 1). In practice, only the parents of some limited set of individuals can be specified. Genealogical relationships are thus defined relative to some set of individuals with unspecified parents; the *founders* of the genealogy. These may be actual immigrants to an isolated population or they may be designated founders in a purely artificial sense. Although this specification of offspring–mother–father triplets is the genealogy, it is of little use as it stands. A useful parametrization of structure must relate to relevant genetic events, such as the survival or ancestry of certain genes, and methods of parametrization must provide methods of data analysis. As an example I shall use the small genealogy of figure 1, which shows useful complex features, but is still easily analysed. Six individuals are assumed to constitute the current population; three of them are supposed to carry a certain type of gene of interest (table 1).

2. Gene identity by descent

Specified genes in a set of individuals are said to be *identical by descent* if all are received by repeated segregation from a single gene in some common ancestor. In this paper, identity of genes will refer always to identity by descent rather than of type. The genes of n specified individuals may be considered as an ordered set of n unordered pairs of genes. The $2n$ genes fall into disjoint subsets, the genes within any subset being identical. However, many of the partitions of the $2n$ genes are genetically equivalent, due to the fact that the two genes within an individual act as an unordered pair in the determination of traits. By defining equivalence relations between partitions obtained from each other from interchanging the two genes of some subset of individuals, one obtains equivalence classes that are genetically distinct *states* of gene identity (Thompson 1974). The number of equivalence classes increases rapidly with n, although not as quickly as the number of partitions. For $n = 6$ there are 4213597 partitions in 198091 genetically distinct gene identity states.

For convenience of example and reference, consider here two summary statistics of the probabilities of gene identity states. The *kinship coefficient*, ψ, between two (not necessarily distinct) individuals B_1 and B_2 is the probability that a gene randomly chosen from B_1 is identical to a gene independently selected from B_2. The *inbreeding coefficient* of an individual is the kinship coefficient between his parents, or the probability that he is *autozygous*; that is, that he carries two identical genes. If the two parents of an individual share no ancestors (relative to the specified genealogy), the individual has zero probability of autozygosity. Between two such individuals there are only three possible states of gene identity: the individuals have i genes in common with probability k_i, $i = 0, 1, 2$ $(k_0 + k_1 + k_2 = 1)$. Their kinship coefficient is

$$\psi(B_1, B_2) = \tfrac{1}{2}\, k_2(B_1, B_2) + \tfrac{1}{4}\, k_1(B_1, B_2), \tag{1}$$

for when the individuals have 1(2) gene(s) in common there is probability $\frac{1}{4}$ $(\frac{1}{2})$ that the gene chosen from each will be identical to each other. More generally, ψ is a linear combination of gene identity state probabilities.

The genealogy of individuals determines a probability of each of the possible states. The converse is not true; state probabilities do not uniquely determine a genealogy. For example, uncle, half-sib and grandparent all have the same state probabilities (table 2). However, the genealogical relationship affects the joint probability distribution of observable genetic traits only through the state probabilities.

$$P(\text{data}\,|\,\text{genealogy}) = \sum_{\text{states}} P(\text{data}\,|\,\text{state})\; P(\text{state}\,|\,\text{genealogy}). \tag{2}$$

Further, any probability statement about types of future joint descendants of the individuals is dependent on the ancestral genealogy only through these current state probabilities. In this precise sense, the genetic consequences of genealogical relationship are summarized by the state probabilities that the genealogy determines.

Table 2. Probabilities of gene identity states between a pair of non-inbred relatives

$k_i = P(i$ genes in common)

	k_2	k_1	k_0	kinship, ψ
parent–offspring	0	1	0	$\frac{1}{4}$
full-sib	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
uncle, half-sib, grandparent	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$
double-first-cousin	$\frac{1}{16}$	$\frac{6}{16}$	$\frac{9}{16}$	$\frac{1}{8}$
quadruple-half-first cousin	$\frac{1}{32}$	$\frac{14}{32}$	$\frac{17}{32}$	$\frac{1}{8}$

Table 3. Probabilities of states of gene identity by descent for individuals 16 and 19 of figure 1

	2^{10} × probability
all four genes identical	1
both autozygous, with distinct genes	3
only 16 autozygous; 1 gene shared with 19	34
only 16 autozygous; 0 genes shared with 19	90
only 19 autozygous; 1 gene shared with 16	10
only 19 autozygous; 0 genes shared with 16	18
neither autozygous; 2 common genes (k_2)	10
neither autozygous; 1 common gene (k_1)	336
neither autozygous; 0 common genes (k_0)	522
total	1024

Although directly related to trait distributions, the set of gene identity state probabilities has two major disadvantages as a parametrization of a genealogy. Not only is there a large number of possible states, but the set of states of positive probability is not easily recognizable from the genealogy. Even between two individuals, each of whom may be autozygous, there are in general 9 states, and, for example, these all have positive probability for the two individuals 16 and 19 of figure 1 (see table 3). On the other hand, only 1794 of the 198091 states between six individuals have non-zero probability for the six current individuals of figure 1. This can only be determined essentially by counting, and a set of 1794 probabilities is in any case an unwieldy specification of their joint relationship.

The other disadvantage is more serious. Genealogical relationships cannot be characterized as probability distributions on the set of gene identity states, for not all distributions are attainable even in the limit. In general the true space of state probabilities is unknown. In the simplest case of relationships between two non-inbred individuals, Thompson (1976) has shown that

$$k_1^2 \geqslant 4k_0k_2. \tag{3}$$

Further, given any specified dyadic-rational k_i $(k_0+k_1+k_2 = 1)$ satisfying (3) a genealogy providing these k_i can be constructed. It is of interest that both restriction and construction derive from a consideration of the cross-parental kinship coefficients. Any dyadic-rational value in [0, 1] is attainable as a kinship coefficient between two individuals in some genealogical structure. Kinship coefficients thus seem to provide a more natural parametrization of relationship. However, they are an insufficient summary of relationship: half-sibs, double-first-cousins and quadruple-half-first-cousins all have the same coefficient of kinship (table 2), but different k_i-values and hence different joint distributions of genetic traits.

3. Descent probabilities

Despite their inadequacies, kinship coefficients are the one universally recognized summary of genealogical structure. They are also readily computed for they satisfy a simple recursive equation. Provided B_1 is not B_2 nor a direct ancestor of B_2,

$$\psi(B_1, B_2) = \tfrac{1}{2}\{(\psi(M_1, B_2)+\psi(F_1, B_2)\}, \tag{4}$$

where M_1 and F_1 are the parents of B_1, since when a gene of B_1 is randomly selected it is a gene of M_1 (F_1) with probability $\frac{1}{2}$ $(\frac{1}{2})$. Further

$$\psi(B_1, B_1) = \tfrac{1}{2}\{1+\psi(M_1, F_1)\}, \tag{5}$$

for the genes chosen 'with replacement' from B_1 are the same gene with probability $\frac{1}{2}$, and are the two distinct genes of B_1 (one from M_1 and one from F_1) also with probability $\frac{1}{2}$.

Karigl (1981) has extended the definition of kinship coefficients to arbitrary numbers of genes. He defines $\psi(B_1, B_2 \ldots, B_n)$ to be the probability that one gene chosen from each of the n individuals, $B_1, B_2 \ldots, B_n$, $(n > 1)$, are all identical. These generalized kinship coefficients satisfy generalizations of (4) and (5), and are related to probabilities of gene identity states by generalizations of (1). Since gene identity state probabilities are uniquely determined by a sufficient set of these generalized multiple kinships, the latter provide an equivalent parametrization of genealogical structure. This parametrization is less directly related to joint distributions of genetic traits (equation (2)), but is more closely related to the original genealogical specification in terms of individuals and their two parents, and to ancestry of genes.

Table 4. Descent probabilities from founder genes to current individuals

(The notation 1(2) denotes that both genes of individual 1 are included in set S, 6(1) that 1 gene of 6 is included, and so on. Probabilities are given as a pair, (i,j), denoting $i/2^j$.)

ancestral set, S	current set				
	{14, 15, 16, 19}	{14, 15, 16}	{16, 19}	{16}	{19}
{1(1)}	17, 14	5, 9	1, 6	1, 3	3, 5
{6(1)}	3, 12	5, 9	3, 8	1, 3	1, 4
{6(2)}	3, 10	3, 7	1, 5	1, 2	1, 3
{6(1), 7(1)}	21, 12	9, 8	5, 7	1, 2	1, 3
{6(1), 1(1)}	61, 13	11, 8	3, 6	1, 2	5, 5
{6(1), 1(2)}	177, 13	43, 9	25, 8	3, 3	1, 2

However, multiple kinships are rather strict in insisting on identity of all of a large number of genes, and rather loose in allowing identity to any ancestral gene. An alternative generalization is to define

$$g_S(B_1, B_2, \ldots, B_n),\ n \geqslant 1,$$

to be the probability that genes chosen from each of the n individuals are all descended from some gene in a specified set of founder genes S (not necessarily from the same gene within this set). Provided B_1 is distinct from individuals $B_2, \ldots, B_n$ (if any) and is not an ancestor of any of them, clearly

$$g_S(B_1, B_2, \ldots, B_n) = (\tfrac{1}{2})\,\{g_S(M_1, B_2, \ldots, B_n) + g_S(F_1, B_2, \ldots, B_n)\}. \tag{6}$$

Further, if $B_1 = \ldots = B_r$ $(1 < r \leqslant n)$ is distinct from and not ancestral to any of the other $(n-r)$ individuals (not themselves necessarily distinct)

$$\begin{aligned} g_S(B_1, B_2, \ldots, B_n) = (\tfrac{1}{2})^{(r-1)}\{&g_S(B_1, B_{r+1}, \ldots, B_n) \\ &+ (2^{(r-1)}-1)\,g_S(M_1, F_1, B_{r+1}, \ldots, B_n)\}, \end{aligned} \tag{7}$$

since the probability that the same gene is selected from B_1 on each of r occasions is $(\frac{1}{2})^{(r-1)}$, and if two different genes are selected they consist of a random gene from each of the parents M_1 and F_1 of B_1. (The functions g_S are, like $\not\psi$, symmetric in their arguments.) These probabilities may thus be computed readily for arbitarily specified founder sets S, g_S satisfying simple boundary conditions where individuals B_i have genes specified to be in S. Computationally, the number n is limited but the number of genes in S is not.

Thus we can compute probabilities that specified current genes descend from various ancestral sets, and, more important, the dependence between descent from certain ancestors to different current individuals. Consider, for example, the genealogy of figure 1. The joint descent probabilities to various of the current individuals from various ancestral sets are given in table 4. Note that descent of a given gene can only increase the probability of descent of the same gene to a relative, and only decrease the probability of descent of other genes. Joint descent to 16 and 19 from the couple (6, 7) has probability $\frac{5}{27}$, whereas the product of the separate probabilities is only $\frac{4}{27}$. On the other hand descent to 16 and 19 from (1, 2) makes descent to 14 and 15 from (6, 7) less probable. In this small genealogy with only four generations interactions are small, but on a large and complex genealogy Thompson (1983) has shown joint descent probabilities more than 100 times the product of separate values.

4. Inferring ancestral types of genes

The descent probabilities of the previous section have direct practical application in inferring the ancestral origins of certain alleles (that is, genes of a certain type) in the current population. We shall denote the particular allele of interest by a_1, and a gene of any other type by a_2. Suppose we have some number of individuals $(B_1, B_2, \ldots, B_n)$ known to carry a_1, and consider a set S of hypothesized ancestral copies of this allele. Then $g_S(B_1, B_2, \ldots, B_n)$ is the probability that a randomly chosen gene in each of these current individuals derives from the ancestral set S, and comparing these probabilities for alternative sets S provides relative likelihoods of these sets as the ancestral allelic a_1 copies.

There are two major oversimplifications here. First, descent only to individuals carrying the a_1 allelle is considered. Any information on its non-descent to other individuals is not included. In figure 1 descent to the assumed carriers (14, 15 and 16) is symmetric between couples (1, 2) and (6, 7) (see table 4) but the fact that 18 and 19 are not carriers must make (6, 7) the more likely founder carriers. Further, analysis of descent only to carriers must bias the analysis towards the conclusion of more ancestral copies. Hypotheses involving different numbers of original copies are not comparable. Secondly, not only are data on current non-carriers disregarded, but also information on types of ancestors. For example, individuals carrying two copies of the a_1 allele may have decreased survival probabilities: often, traits of interest in large and complex genealogies are of this recessive type. Ancestors then have some reduced (perhaps even zero) probability of having carried two such alleles. In a complex genealogy over many generations inclusion of this fact can alter inferences.

Against these disadvantages there are two advantages. Although for simplicity the genes of S were above specified as being genes of founders, in fact, provided S does not involve individuals who are ancestors of each other, they may be any ancestral genes. Hence descent of a particular allele may be traced down the genealogy, by hypothesizing ancestral sets S at different generations. The second advantage is the ease of computation: many alternative hypotheses may be very rapidly assessed. These advantages are apparent in a re-analysis of the data of Kidd *et al.* (1980) on the ancestry of propionic acidaemia in a Mennonite–Amish genealogy. The two disadvantages also apply, but not with strong force, since individuals with two copies of the allele can be without clinical symptoms, and few individuals among the ancestors have *a priori* high probability of carrying two genes identical by descent. Thompson (1983) shows that patterns of joint descent, jointly between current carrier parents of affected individuals and jointly between alternative founder carriers, are important in a quantitative assessment of the possible hypotheses.

If the above is an oversimplified approximation, what is the full solution? Suppose that, for a given combination of types of original founder genes, one could compute the probability of the data observed on current individuals of all types, under a given genetic model, perhaps involving information about varying viability of types of ancestors. This would then be a likelihood for that combination of founder types. In principle this can be done, using the method of Cannings *et al.* (1978), the basis of which is the following. Define a *cutset* of individuals to be a set who together divide the genealogy. For present purposes it will be sufficient to consider cutsets dividing a current set of individuals from a set of ancestors including all the founders of the genealogy; for example, individuals {12, 13, 18, 10, 11} in figure 1. If the types of the genes carried by such a set of individuals are specified, genetic events in sets of individuals on different sides of the cutset are statistically independent. Data on individual 4, for example, convey no information about the traits of 12 and 15, and vice versa. We then consider probabilities of data below a given cutset, conditional on each possible combination of types of genes in individuals in the cutset, and work sequentially back through the genealogy from one cutset to the next, incorporating parent–offspring segregation probabilities and any other information on traits of individuals of types of ancestral genes. Finally we obtain the probability of all data observed on the genealogy given each possible (ordered) combination of founder gene types, or simultaneously the likelihoods of every possible founder combination.

Table 5. Part of the ancestral likelihood for the pedigree of figure 1 under the data of table 1

(Carriers are known to carry one a_1 gene, the other current individuals none. Founders 3 and 17 are here taken as the most likely combination a_2a_2, and the marginal likelihood for the other two founder couples is tabulated, the full function being given by symmetry between the two members of each couple. Figures in brackets give the likelihood when no ancestor can have carried two a_1 genes. The numbers each divided by 2^{15} give the exact probability of data under the ancestral combination.)

couple (1, 2)	couple (6, 7)					
	$a_1a_1 \times a_1a_1$	$a_1a_1 \times a_1a_2$	$a_1a_1 \times a_2a_2$	$a_1a_2 \times a_1a_2$	$a_1a_2 \times a_2a_2$	$a_2a_2 \times a_2a_2$
$a_1a_1 \times a_1a_1$	0	0	0	0	0	0
$a_1a_1 \times a_1a_2$	0	60	160	250	560	1200
$a_1a_1 \times a_2a_2$	0	192	384	480	768	1152
$a_1a_2 \times a_1a_2$	0	300	480	720 (240)	1140 (475)	2016 (560)
$a_1a_2 \times a_2a_2$	0	784	896	1330 (665)	1260 (1260)	1232 (1232)
$a_2a_2 \times a_2a_2$	0	1920	1536	2688 (896)	1408 (1408)	0 (0)

In table 5 is shown part of the ancestral likelihood function for the example genealogy. Note that couple (6, 7) are indeed the more likely ancestral a_1 carriers, it being most likely that both members of the couple are so. Note also the ordering of the likelihoods, which may be unexpected *a priori*. The different orderings between rows and between columns show the necessity for joint inferences on the two couples, even in this small example. The effect of excluding possible a_1a_1 ancestors is here to reduce the likelihood of $a_1a_2 \times a_1a_2$ founder couples; not surprisingly, since the other ancestors of the current population consist mainly of offspring of these couples. In a larger genealogy the decreased likelihood of any such ancestral couples can have varied effects on inferences about original founders.

In a large complex genealogy, this approach may not be computationally feasible. At each stage all possible combinations of types of genes for all members of the current cutset must be considered. Although it is sometimes possible to work sequentially through large genealogies of isolated populations with cutsets of no more than 9 or 10 individuals, this is not always feasible and determining the optimal cutset sequence is in general an unsolved problem. Further, the number of founders may be prohibitive. In many cases it will be necessary to consider jointly only some subset of the founders, under some (probabilistic) assumptions about the types of the remainder. One population for which this can be done is the small isolated population of Tristan da Cunha, where eleven early founders contribute 80 % of current genes. Here Thompson (1978) has shown that inferences can be made about the joint types of genes in founders living before 1827. The multiple complex paths of relationship increase computational difficulty, but provide the information required. Such inferences are limited to simple genetic traits. Nonetheless, the power to make inferences about the types of genes seven generations ago indicates that, conversely, these types can affect current trait distributions. This example of the Tristan da Cunha population is considered further below.

5. Gene survival and gene extinction

This analysis of ancestry in terms of joint likelihoods on sets of founders leads to an alternative characterization of genealogical structure. For the essence of evolution in a population is gene survival: the number and variety of distinct surviving genes. So instead of ancestry let us consider gene survival, or equivalently extinction. Just as descent probabilities have interpretation as approximate ancestral likelihoods, so the complete ancestral likelihoods provide extinction probabilities. Consider a trait for which there are just two alleles a_1 and a_2, and a specified combination of alleles among original founders. Then the probability of extinction of (at least) those founder genes labelled a_1 is the probability that, given the ancestral combination, the population now consists entirely of individuals with two a_2 genes. But this is also the ancestral likelihood of the particular combination of ancestral a_1 and a_2 genes, given this current population. Working backwards from a current population in which all individuals are assumed to carry two a_2 genes, we can therefore compute simultaneously the extinction probabilities of all combinations of founder genes.

Again a joint analysis is important. Survival of the genes of a given founder over a specified genealogy decreases the survival probabilities of genes in other founder individuals who share descendants with the first. Particularly, therefore, survival decreases survival of spouse genes, and, indeed, survival of one gene in a founder decreases the survival probability of the other. Similarly extinction of some genes decreases extinction probabilities of others: some genes must survive in an extant population. The six individuals of figure 1 carry at least three and at most nine distinct genes, although there are six founders to the genealogy. Not all four genes of either original couple can be extinct, nor both those of 17. Although there is little interaction between the two couples, since the population consists mainly of their grandchildren, survival of a gene of 3 decreases the probability of survival of all four genes of (1, 2) from $\frac{9}{2^7}$ to $\frac{1}{2^5}$. Survival of both genes of 7 decreases the survival probability of both genes of 6 from $\frac{15}{2^6}$ to $\frac{2}{15}$. The extent to which survival or extinction of certain subsets of founder genes precludes survival or extinction of other disjoint subsets provides a measure of the structure of the genealogy with respect to the limited paths for descent of genes.

How many genes do survive in a small isolated population? Questions about the exact numbers of genes are not precisely the same as those about the fate of (at least) a certain labelled set of genes. However, answers to the latter, which are provided by the ancestral likelihoods, can be transformed to provide the required probability distributions. To turn finally to a real example again, the eleven early founders of the Tristan de Cunha population provided 22 potential genes, but not all can be present now. Thomas & Thompson (1983) have shown that with probability 0.994 between 10 and 18 genes survive, these being made up of between 4 and 6 of the six genes in the three founder females and of between 6 and 13 of the sixteen genes in the eight founder males. Although interactions are generally small in this expanding population, survival of some founder genes does reduce survival probabilities for others; note that 6 (female genes) + 13 (male genes) > 18 (total genes). Such analyses emphasize just how rapid the loss of variability can be in a small isolated population, and just how crucial certain segregations can be in determining the current genetic constitution.

References

Cannings, C., Thompson, E. A. & Skolnick, M. H. 1978 Probability functions on complex pedigrees. *Adv. appl. Prob.* **10**, 26–61.

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

Kidd, J. R., Wolf, B., Hsia, Y. E. & Kidd, K. K. 1980 Genetics of propionic acidemia in a Mennonite-Amish kindred. *Am. J. hum. Genet.* **32**, 236–245.

Thomas, A. & Thompson, E. A. 1983 The number of genes on Tristan da Cunha. (Submitted.)

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

Thompson, E. A. 1976 A restriction on the space of genetic relationships. *Ann. hum. Genet.* **40**, 201–204.

Thompson, E. A. 1978 Ancestral inference. II. The founders of Tristan da Cunha. *Ann. hum. Genet.* **42**, 239–253.

Thompson, E. A. 1983 A recursive algorithm for inferring gene origins. *Ann. hum. Genet.* **47**, 143–152.

Discussion

A. W. F. Edwards (*Caius College, Cambridge University, U.K.*). How does one prove that every dyadic ratio for a kinship coefficient corresponds to some genealogical relationship?

Elizabeth A. Thompson. The form of equations (4) and (5) shows that this must be so, and the fact has been known for a very long time. However, the nicest proof I know is a constructive demonstration given only recently by Dr G. Karigl. Expressing the required kinship coefficient as a binary expansion, the sequence of zeros and ones can be used to define an explicit sequence of outbreeding and backcrossing which produces the required result. Although such a genealogy is unlikely in human populations, this neatly proves the theoretical result.

At the meeting, Professor Felsenstein, Professor Hill, Professor Bodmer and Professor Kingman raised questions of complexity of genealogies, inaccuracies in genealogies, complex genetic models and the approximations it is then necessary to introduce into computations. In principle, the methods of obtaining ancestral likelihoods apply to arbitrarily complex genetic models on arbitrarily complex genealogies. In practice, there are of course computational limitations, although really quite complex situations can be considered. In my paper I have covered what might be referred to as 'the theory of exact computations on genealogies'. The next stage, which requires both theoretical and practical work, is a study of approximate computations. By how much are results altered by omitting apparently uninformative sections of genealogy? How dependent are results on certain critical links in a genealogy, and how can we determine which they are? What is the expected gain in using linked loci to increase the power to make inferences? How much is lost by having only phenotypic rather than genotypic data? Although some work has been done in this area, these remain important open questions.