J. Math. Biol. (1988) 26: 57-64

Journal of
Mathematical
Biology

Calculation of the inbreeding coefficient

Wayne Boucher*,**

Enrico Fermi Institute, University of Chicago, Chicago, IL 60637, USA

Abstract. Wright's rule for calculating the inbreeding coefficient for an arbitrary pedigree is proven for both autosomal and X-linked loci.

Key words: Inbreeding — Graph theory

Inbreeding increases the chance that offspring will inherit two copies of a harmful recessive gene, and so is not a good genetic strategy. A measure of the inbreeding of the ancestors of an individual, I, is provided by the inbreeding coefficient (Wright, 1922), F_I, which is defined to be the probability that the two genes at a given locus of I are identical by descent (Cotterman 1940; Malécot 1941, 1942, 1948), that is, either they are derived from the same gene, or one is derived from the other. More generally, a measure of the relatedness of two individuals, A and B, is provided by the coefficient of consanguinity (Malécot 1941, 1942, 1948), F_{AB}, which is the probability that a randomly chosen gene from A and a randomly chosen homologous gene from B are identical by descent.

Given a model of population genetics it is possible in principle to calculate F_I and F_{AB}. In this paper it will be assumed that mating is independent of genotype and that there is no selection or mutation.

Initially consider autosomal loci. In this case, the sexes need not and will not be distinguished. With these assumptions, if I is a (genuine or hypothetical) offspring of A and B, then $F_{AB} = F_I$, and so only the inbreeding coefficients need to be calculated. Also, F_I is a function only of ancestry, and is the same for all loci and independent of gene frequencies. Given the ancestry of I, Wright (1922) gave a simple rule to calculate F_I. Malécot (1941) offered a proof of this rule, but it is not clear that his proof covers all possible cases. Here, a new proof will be given, by proving a slightly more general result. Although the idea behind the proof is rather simple (it is basically an algorithm that works from generation to generation) there is unfortunately a fair amount of notation that needs to be introduced.

The family tree of the individual I can be represented as a (possibly non-planar) oriented graph $G \subseteq \mathbb{R}^2$ (Bollobás 1979), constructed so that the vertices,

* Supported by NSF Grant PHY-84-16691

** *Current address*: Trinity College, Cambridge CB2 1TQ, England

V, of G represent I and his ancestors, and so that the edges, E, represent parenthood, the direction of the edges being from the offspring to the parents. G can be non-planar in that the edges can cross without a vertex at the intersection. V and E will be identified with their images in $\mathbb{R}^2$. The edges will be considered to be open intervals (and so not to include their endpoints).

Note, there is more structure implied in a family tree than in an arbitrary oriented graph, in that two edges exit from every vertex (i.e. every individual has two parents), no edges enter the vertex corresponding to I (i.e. the descendants of I are not considered), no more than one edge can connect two vertices (i.e. there is no selfing), and there can be no uni-directional loops (i.e. if A is an ancestor of B, then B is not an ancestor of A). The fact that G is embedded in $\mathbb{R}^2$ is not a restriction on the pedigree because G can be non-planar, and the embedding is used only to make the results easier to state and prove. G is not necessarily a finite graph.

It is convenient to introduce a map $h: V \to \mathbb{N}$ to represent the generations, h being arbitrary except that $h(I) = 1$ and if B is an ancestor of A then $h(B) > h(A)$ (so that, in particular, the generations are counted backwards starting from I). That such a map exists is a mild assumption on the pedigree, and will always be true in biologically realistic cases. An example of a pedigree where such a map does not exist is the mating of a man with his daughter, granddaughter, etc. Given that an h exists, one way to construct one is to let $h(A)$ equal the maximum length (plus one) of all uni-directional paths from I to A, but it is not necessary to use this particular h. In any case, each generation, n, has a finite number ($\leqslant 2^{n-1}$) of individuals in it.

The graph, G, is arranged in $\mathbb{R}^2$ as follows. First, the vertices are placed so that $A \in \mathbb{R} \times \{h(A)\} \subseteq \mathbb{R} \times \mathbb{N}$ for all $A \in V$ and so that no two vertices are placed on top of each other. Then the edges are drawn so that if A is an offspring of B, then the corresponding edge lies in $\mathbb{R} \times (h(A), h(B))$. Further, by perturbing the edges if necessary, it is arranged so that two or more edges do not intersect in $\mathbb{R} \times \mathbb{N}$ and no vertex lies on (but only at the end of) an edge. Of course, since $V \subseteq \mathbb{R} \times \mathbb{N}$, two or more edges can be incident at a point (vertex) in $\mathbb{R} \times \mathbb{N}$. (The problem of edges intersecting in $\mathbb{R} \times \mathbb{N}$ arises only because of the possible non-planarity of G.)

For accounting purposes, it is convenient to introduce a new graph, $G' \subseteq \mathbb{R}^2$, constructed as follows. Let $V'_i = G \cap (\mathbb{R} \times \{i\})$, $i \in \mathbb{N}$. As a set, this consists of one point for each vertex $A \in V$ such that $h(A) = i$, and also one point for each edge of G connecting two vertices B, $C \in V$ such that $h(B) < i$ and $h(C) > i$.

Let G' be the oriented (possibly non-planar) graph with vertices $V' = \bigcup V'_i$, and, with (directed) edges, E', induced from those of G. Call those vertices in G' which correspond to vertices in G "individuals" and call the others "virtual points" (in the figures, circles will correspond to the former and points to the latter). The elements of V'_i will be called generation i. Let V_i be the individuals in V'_i. Then $V = \bigcup V_i$. Let E'_i be the set of edges in G' between generations i and $i+1$. As a set in $\mathbb{R}^2$, $E'_i = G' \cap (\mathbb{R} \times (i, i+1))$. An example of this construction is given in Fig. 1.

For an edge $a \in E'_i$, let $\mu(a) \in V'_i$ be the vertex at which a starts and let $\nu(a) \in V'_{i+1}$ be the vertex at which a ends. It is convenient to identify a with the

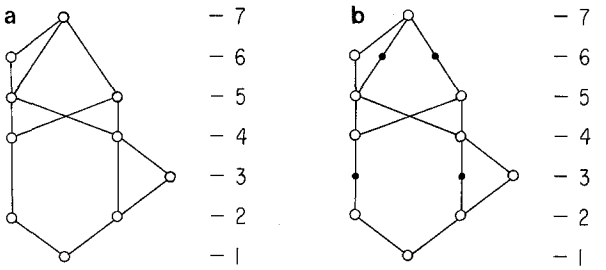

Fig. 1. Example of the graph, G, of a family tree, and of the corresponding graph, G'. (These graphs are non-planar.) The individual, I, is at the bottom of each graph and the generations are numbered at the right. Only ancestors which are needed to calculate the inbreeding coefficient are shown

gene donated to $\mu(a)$ by $\nu(a)$. Then, if $a, b \in E_i'$, let $P(a \equiv b)$ be the probability that a and b are identitical by descent.

For $A \in V$, let F_A be the inbreeding coefficient of A. The rule for the calculation of F_I given in the theorem below will be proven by induction and will involve relating probabilities of identity by descent in one generation to those in the next. For $a, b \in E_i'$, $a \neq b$, there exist the following three possibilities in generation $i+1$:

(i) $\nu(a), \nu(b) \in V_{i+1}$. See Fig. 2a for notation. Then, since there is a probability of $\frac{1}{2}$ that a randomly chosen gene has come from each parent,

$$P(a \equiv b) = \tfrac{1}{2}(P(a \equiv e) + P(a \equiv f))$$
$$= \tfrac{1}{4}(P(c \equiv e) + P(c \equiv f) + P(d \equiv e) + P(d \equiv f)). \quad (1)$$

A special case of this is if $\nu(a) = \nu(b)$ (see Fig. 2b), in which case

$$P(a \equiv b) = \tfrac{1}{2}(1 + P(c \equiv d)) = \tfrac{1}{2}(1 + F_{\nu(a)}). \quad (2)$$

(ii) $\nu(a) \notin V_{i+1}$, $\nu(b) \in V_{i+1}$. See Fig. 2c.
Then,

$$P(a \equiv b) = \tfrac{1}{2}(P(c \equiv d) + P(c \equiv e)). \quad (3)$$

(Note, in G, $\nu(a)$ is not an individual, and so a and c are really the same gene.) Of course, a similar formula can be given if $\nu(a) \in V_{i+1}$, $\nu(b) \notin V_{i+1}$.

(iii) $\nu(a), \nu(b) \notin V_{i+1}$. See Fig. 2d.
Then

$$P(a \equiv b) = P(c \equiv d). \quad (4)$$

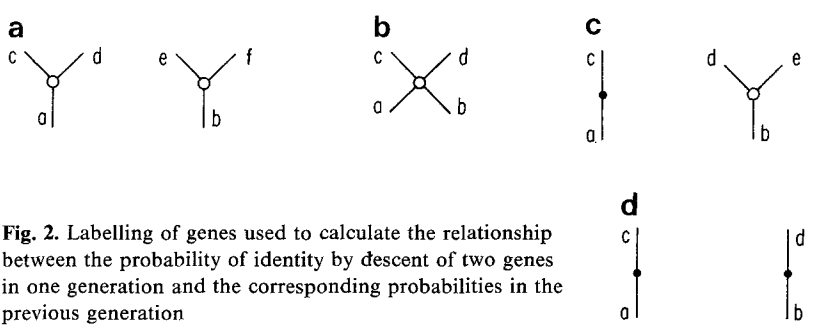

Fig. 2. Labelling of genes used to calculate the relationship between the probability of identity by descent of two genes in one generation and the corresponding probabilities in the previous generation

(Note, as above, a and c are the same gene and similarly b and d are the same gene.)

For $a \in E_i'$, $b \in E_{i+1}'$, let

$$R_{a,b} = \begin{cases} 1 & \text{if } \nu(a) = \mu(b), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

For $a, a' \in E_i'$, let

$$w_{a,a'} = \begin{cases} 2 & \text{if } \nu(a), \nu(a') \text{ are both individuals,} \\ 1 & \text{if exactly one of } \nu(a), \nu(a') \text{ is an individual,} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

For $a, a' \in E_i'$, let

$$\delta(\nu(a) = \nu(a')) = \begin{cases} 1 & \text{if } \nu(a) = \nu(a') \in V_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

For $a, a' \in E_i'$, let

$$\delta(\nu(a) \neq \nu(a')) = \begin{cases} 1 & \text{if } \nu(a) \neq \nu(a'), \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The rules (i)-(iii) given above imply that if $a, a' \in E_i'$, $a \neq a'$, then

$$P(a \equiv a') = \left(\sum_{b,b' \in E_{i+1}'} R_{a,b} R_{a'b'} P(b \equiv b') \right) \left(\tfrac{1}{2}\right)^{w_{a,a'}} \delta(\nu(a) \neq \nu(a')) + \tfrac{1}{2}(1 + F_{\nu(a)}) \delta(\nu(a) = \nu(a')). \tag{9}$$

For $a \in E_i'$, a (uni-directional) path, z_a, (starting from I) is a sequence of edges $z_a = (a_1, \ldots, a_i = a)$ such that $a_j \in E_j'$ and $\nu(a_j) = \mu(a_{j+1})$ for all $1 \leq j \leq i-1$ (in particular, $a_1 \in E_1'$ implies that $\mu(a_1) = I$). Two such paths $z_a = (a_1, \ldots, a_i = a)$ and $z_{a'} = (a_1', \ldots, a_i' = a')$ are said to be disjoint if $a_j \neq a_j'$ for all $1 \leq j \leq i$. Note that this does not preclude $\nu(a) = \nu(a')$ (and we always have $\mu(a_1) = \mu(a_1') = I$).

For $a, a' \in E_i'$, let $X_{a,a'}$ be the set of ordered pairs of paths $x = (z_a, z_{a'})$ such that z_a and $z_{a'}$ are disjoint. For $x \in X_{a,a'}$ such a pair of paths, let n_x be the number of individuals (i.e. vertices that are not virtual points) crossed by the two paths, taken together (but not counting I), or, more formally, the number of individuals in the set $\{\nu(a_1), \ldots, \nu(a_{i-1}), \nu(a_1'), \ldots, \nu(a_{i-1}')\}$. If $X_{a,a'} \neq \varnothing$, let

$$f_{a,a'} = \tfrac{1}{2} \sum_{x \in X_{a,a'}} \left(\tfrac{1}{2}\right)^{n_x} \tag{10}$$

(the factor of $\frac{1}{2}$ in front arises because of double-counting later), and let $f_{a,a'} = 0$ if $X_{a,a'} = \varnothing$. Note in particular that $f_{a,a'} = 0$ if $\mu(a) = \mu(a')$ (except that $f_{a,a'} = \frac{1}{2}$ if $\mu(a) = \mu(a') = I$, $a \neq a'$).

For $A \in V_{i+1}$ (i.e., A is an individual, not a virtual point, in generation $i+1$) let Y_A be the set of ordered pairs of paths $y = (z_a, z_{a'})$ such that z_a and $z_{a'}$ are disjoint and $\nu(a) = \nu(a') = A$ (taking $Y_I = \varnothing$ by convention). For $y \in Y_A$ such a pair of paths, let n_y be the number of individuals crossed by the two paths, taken together (counting A, but not counting I), or, more formally, the number of

individuals in the set $\{\nu(a_1), \ldots, \nu(a_{i-1}), \nu(a'_1), \ldots, \nu(a'_{i-1})\}$ *plus one* (the extra one is because A is counted). If $Y_A \neq \varnothing$, let

$$g_A = \tfrac{1}{2} \sum_{y \in Y_A} (\tfrac{1}{2})^{n_y}, \tag{11}$$

(the factor of $\frac{1}{2}$ in front arises because we have taken Y_A to be the set of ordered, not unordered, pairs of paths) and let $g_A = 0$ if $Y_A = \varnothing$. Note in particular that $g_I = 0$.

For every pair of paths $y \in Y_A$, $A \in V_{i+1}$, there are two unique ordered edges $a, a' \in E'_i$ with $\nu(a) = \nu(a') = A$, and a unique pair of paths $x \in X_{a,a'}$ with $y = x$. Conversely, given $a, a' \in E'_i$ such that $\nu(a) = \nu(a') = A$, then for every pair of paths $x \in X_{a,a'}$ there is a unique pair of paths $y \in Y_A$ with $y = x$. In this correspondence, we have that $n_y = n_x + 1$ (remember, A is counted in n_y). Thus, for all $A \in V_{i+1}$,

$$g_A = \tfrac{1}{2} \sum_{a,a' \in E'_i, \nu(a) = \nu(a') = A} f_{a,a'}. \tag{12}$$

Note, the double-counting mentioned above arises here because in this sum a and a' are taken to be an ordered pair, and so $f_{a,a'}$ and $f_{a',a} = f_{a,a'}$ both contribute (also, $f_{a,a} = 0$ for all a). The factor of $\frac{1}{2}$ in front arises because $n_y = n_x + 1$.

Hereafter, a, a' will be elements of E'_i and b, b' will be elements of E'_{i+1}, $i \geqslant 1$. If $\mu(b) = \mu(b')$, then $f_{b,b'} = 0$ (since $i+1 \geqslant 2$ and so $\mu(b) \neq I$). If $\mu(b) \neq \mu(b')$ then (using reasoning similar to that used in deriving Eq. (12)),

$$\begin{aligned} f_{b,b'} &= \tfrac{1}{2} \sum_{x \in X_{b,b'}} (\tfrac{1}{2})^{n_x} \\ &= \tfrac{1}{2} \sum_{a,a' \in E'_i} \sum_{\bar{x} \in X_{a,a'}} (\tfrac{1}{2})^{w_{a,a'}} (\tfrac{1}{2})^{n_{\bar{x}}} R_{a,b} R_{a',b'} \\ &= \sum_{a,a' \in E'_i} (\tfrac{1}{2})^{w_{a,a'}} f_{a,a'} R_{a,b} R_{a',b'}. \end{aligned} \tag{13}$$

(If $X_{b,b'} = \varnothing$, so that $f_{b,b'} = 0$, then the last sum is also zero.)

If $\mu(b) \neq \mu(b')$, then the non-zero contributions in the last sum satisfy $\delta(\nu(a) \neq \nu(a')) = 1$, because $\nu(a) = \mu(b) \neq \mu(b') = \nu(a')$ if $R_{a,b} = R_{a',b'} = 1$. Thus, it is possible to rewrite this equation as

$$f_{b,b'} = \sum_{a,a' \in E'_i} (\tfrac{1}{2})^{w_{a,a'}} f_{a,a'} R_{a,b} R_{a',b'} \delta(\nu(a) \neq \nu(a')), \tag{14}$$

and now this is also true when $\mu(b) = \mu(b')$.

Theorem. *With the notation introduced above, for all* $i \geqslant 1$,

$$F_I = \sum_{j=1}^{i} \left(\sum_{A \in V_j} (1 + F_A) g_A \right) + \sum_{a,a' \in E'_i} P(a \equiv a') f_{a,a'}. \tag{15}$$

(*Note that there is double-counting in the second sum, for the reason given previously.*)

Proof. The formula is proven by induction on i. It is true for $i = 1$, since it then reduces to $F_I = P(a_1 \equiv b_1)$, where a_1 and b_1 are the genes donated to I by his

parents. Assume the formula is true for some $i \geqslant 1$. Then

$$\sum_{a,a' \in E_i'} P(a \equiv a') f_{a,a'} = \sum_{a,a' \in E_i'} \Bigg[\sum_{b,b' \in E_{i+1}'} R_{a,b} R_{a',b'} P(b \equiv b') \left(\tfrac{1}{2}\right)^{w_{a,a'}} \delta(\nu(a) \neq \nu(a')) + \tfrac{1}{2}(1 + F_{\nu(a)}) \delta(\nu(a) = \nu(a')) \Bigg] f_{a,a'}$$

$$= \sum_{b,b' \in E_{i+1}'} P(b \equiv b') f_{b,b'} + \sum_{A \in V_{i+1}} (1 + F_A) g_A, \tag{16}$$

using Eq. (9) and then Eqs. (14) and (12).

Therefore,

$$F_I = \sum_{j=1}^{i} \left(\sum_{A \in V_j} (1 + F_A) g_A \right) + \sum_{a,a' \in E_i'} P(a \equiv a') f_{a,a'}$$

$$= \sum_{j=1}^{i+1} \left(\sum_{A \in V_j} (1 + F_A) g_A \right) + \sum_{b,b' \in E_{i+1}'} P(b \equiv b') f_{b,b'}, \tag{17}$$

which finishes the induction.

Corollary. *With the notation as introduced above, if there is some generation i such that for all* $a, a' \in E_i'$ *either* $X_{a,a'} = \varnothing$ *or* $P(a \equiv a') = 0$ *then*

$$F_I = \sum_{j=1}^{i} \left(\sum_{A \in V_j} (1 + F_A) g_A \right)$$

$$= \sum_{j=1}^{i} \left(\sum_{A \in V_j} (1 + F_A) \tfrac{1}{2} \sum_{y \in Y_A} \left(\tfrac{1}{2}\right)^{n_y} \right), \tag{18}$$

which is the rule given by Wright (1922).

The theorem and corollary can be shown to be extendable to the case when the locus is on the X chromosome, if the following modifications are made (no proofs will be given because they are similar to the autosomal case). The rule was stated by Wright (1933); see also Crow and Kimura (1970), p. 73.

In the graph G, the vertices now split into females and males, and it is necessary to erase all edges that both start and end at a male. The graph G' is then constructed as before, except that the individuals now split into females and males.

Males have only one copy of the X chromosome, and so the inbreeding coefficient is not defined for them. However, it is convenient to define $F_A = 0$ if A is a male. Also, let

$$\alpha_A = \begin{cases} 1 & \text{if } A \text{ is female,} \\ 2 & \text{if } A \text{ is male.} \end{cases} \tag{19}$$

The rules (i)-(iii), as given by Eqs. (1)-(4), have to be modified to distinguish between male and female vertices. Rather than give the new rules, it is sufficient to note that Eq. (9) will remain valid if Eqs. (5) and (8) are left unchanged and

if the following changes are made in Eqs. (6) and (7):

$$w_{a,a'} = \begin{cases} 2 & \text{if } \nu(a),\ \nu(a') \text{ are both female,} \\ 1 & \text{if exactly one of } \nu(a),\ \nu(a') \text{ is female,} \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

$$\delta(\nu(a) = \nu(a')) = \begin{cases} \alpha_{\nu(a)} & \text{if } \nu(a) = \nu(a') \in V_{i+1}, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Equations (10) and (11) should be left unchanged, except that n_x and n_y are now the number of females crossed by the two paths. However, Eq. (12) now becomes

$$g_A = \tfrac{1}{2}\alpha_A \sum_{a,a' \in E'_i, \nu(a) = \nu(a') = A} f_{a,a'}, \tag{22}$$

where the factor of α_A arises because $n_y = n_x$ if A is male, whereas $n_y = n_x + 1$ (as before) if A is female.

With these changes, the remaining equations (13)–(18) remain valid, and in particular the theorem and corollary are true.

Probabilistic interpretation. P. Arzberger has pointed out to the author (in a referee's report) that the results given here may also be obtained by the introduction of two independent, identically distributed Markov chains $\{X_i\}$ and $\{Y_i\}$ on G'. For example, for autosomal loci the transitions for the X_i are determined as follows (those for the Y_i are given similarly): for $a \in E'_i$, $b \in E'_{i+1}$ let

$$P[X_{i+1} = b, X_i = a] = \begin{cases} 1 & \text{if } \nu(a) = \mu(b) \in V_i, \\ \frac{1}{2} & \text{if } \nu(a) = \mu(b) \notin V_i, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

(It is assumed that $X_i,\ Y_i \in E'_i$ for all i, and so, in particular, $\mu(X_1) = \mu(Y_1) = I$.) Then, for example, if $a,\ a' \in E'_i$,

$$f_{a,a'} = P[X_i = a,\ Y_i = a',\ X_j \neq Y_j,\ 1 \leqslant j \leqslant i], \tag{24}$$

and, if $A \in V_{i+1}$,

$$g_A = P[\nu(X_i) = \nu(Y_i) = A,\ X_j \neq Y_j,\ 1 \leqslant j \leqslant i]. \tag{25}$$

All of the formalism may be rewritten in terms of the X_i and Y_i.

Acknowledgements. I would like to thank T. Nagylaki for suggesting the problem and reading the paper, and P. Arzberger and an anonymous referee for suggesting improvements in the presentation of the paper, in particular with regard to the graph theory and the definition of the generation number.

References

Bollobás, B.: Graph theory. Berlin Heidelberg New York: Springer 1979

Cotterman, C. W.: A calculus for statistico-genetics. Thesis, Ohio State University, Columbus (1940). Reprinted in: Balloutti, P. (ed.) Genetics and social structure. Stroudsburg, Pennsylvania 1974

Crow, J. F., Kimura, M.: An introduction to population genetics theory. New York: Harper and Row 1970

Malécot, G.: Etude mathématique des populations "mendéliennes". Ann. Univ. Lyon, Sciences **A4**, 45-60 (1941)

Malécot, G.: Mendélisme et consanguinité. Comptes Rendus Acad. Sci. Paris **215**, 313-314 (1942)

Malécot, G.: Les mathématiques de l'hérédité. Paris: Masson 1948. Extended translation: The mathematics of heredity. San Francisco: Freeman 1969

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

Wright, S.: Inbreeding and homozygosis. Proc. Natl. Acad. Sci. **29**, 411-420 (1933)

Received July 28/Revised September 26, 1987