Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 240 (2006) 360–367
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Reconstructing pedigrees: A combinatorial perspective

Mike Steel[a,*], Jotun Hein[b]

[a]Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, Christchurch, New Zealand
[b]Department of Statistics, University of Oxford, UK

Received 19 August 2005; accepted 27 September 2005
Available online 1 December 2005

Abstract

A pedigree is a directed graph that displays the relationship between individuals according to their parentage. We derive a combinatorial result that shows how any pedigree—up to individuals who have no extant (present-day) ancestors—can be reconstructed from (sex-labelled) pedigrees that describe the ancestry of single extant individuals and pairs of extant individuals. Furthermore, this reconstruction can be done in polynomial time. We also provide an example to show that the corresponding reconstruction result does not hold for pedigrees that are not sex-labelled. We then show how any pedigree can also be reconstructed from two functions that just describe certain circuits in the pedigree. Finally, we obtain an enumeration result for pedigrees that is relevant to the question of how many segregating sites are needed to reconstruct pedigrees.

Keywords: Pedigree; Isomorphism; Reconstruction; Enumeration

1. Introduction

The reconstruction of ancestral relationships—between genera, species and populations—is central to much of evolutionary biology. Traditionally, phylogenetic methods have been used at the species (and higher) level, while for populations, graph-based approaches (such as haplotype networks) have been more commonly employed. However although phylogeny dates back to Darwin, people have been interested for much longer in the more basic level of ancestral—namely that of pedigrees, which traces the ancestors of individuals back in time.

The reconstruction of pedigrees—driven by human curiosity, or more practical requirements (legal, medical, etc.)—has variously involved meticulous record-taking over centuries, the methodical recording and searching of birth-and-death records, oral tradition, and verbal transmission of ancestry (as with the Maori *whakapapa*). Of course this mixture of approaches, along with a great deal of 'missing data' leads to a very patchy and incomplete pedigree for most collections of extant individuals. A detailed (complete) record of ancestry would typically run only a few generations back before ambiguities arise. For example, it would be rare for two individuals chosen at random in a major continent to be able to tell how many ancestors they shared (say) five generations ago. Similarly although nearly all individuals have four distinct grandparents, few of us could confidently state how many *distinct* ancestors we had exactly 10 generations ago (some integer in the range from 2 to 1024). However, it seems likely that the increasing availability of genomic data for individuals (rather than species) will soon provide new data and impetus for reconstructing pedigrees, though how far into the past this might be possible is currently unclear.

To date, the formal analysis of pedigrees has concentrated much less on reconstruction, and more on carrying out statistical tests on the genetic aspects of pedigrees (as for example, in the pioneering work of Cannings and Thompson (1981) and Thompson (2000)). In other areas of systematic biology—for example in the reconstruction of phylogenetic trees—two questions arise: firstly, can we piece together parts of a tree of life from smaller subtrees (Bininda-Emonds, 2004)? Secondly, how much data (DNA sequences) would we need to reconstruct such a large tree

*Corresponding author. Tel.: +64 3 366 7001; fax: +64 3 364 2587.
E-mail addresses: m.steel@math.canterbury.ac.nz (M. Steel), hein@stats.ox.ac.uk (J. Hein).

0022-5193/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2005.09.026

(Mossel and Steel, 2005; Sober and Steel, 2002)? Here we ask the analogous questions for pedigrees—can they be reconstructed formally from sub-pedigrees? And if we use sequences to reconstruct pedigrees, how many segregating sites would be needed given that the number of pedigrees is likely to grow very quickly? In this paper, we provide some initial answers to these questions using basic combinatorial arguments. However, we also see considerable scope for future work on the questions we posed above, and we list some particular problems at the end of this paper. Moreover, our results are not intended to be tailor-made computational techniques for application to specific data; rather we describe a general mathematical approach as a framework for investigating pedigree reconstruction questions. We hope to explore further how these ideas might be applied for the reconstruction of pedigrees from genomic data. We begin with some formal definitions, which follow (Semple and Steel, 2003; Thomas, 1993).

1.1. Definitions

A (*strict*) *pedigree* $\mathscr{P}$ is an acyclic digraph, for which the vertex set V is the disjoint union of two subsets M and F ('Male' and 'Female') and for which each vertex $v \in V$ satisfies the following condition:

(P) if v has positive indegree then v has exactly two incoming arcs, say (u, v) and (u', v), where $u \in M$ and $u' \in F$.

Condition (P) formalizes the requirement that each individual in the set V that has at least one parent in V has exactly two, one male and one female.

We will write $\mathscr{P} = (V, A)$ to indicate that pedigree $\mathscr{P}$ has vertex set V and arc set A. If X is a subset of the vertices of $\mathscr{P}$ that have no outgoing arcs, then we say that $\mathscr{P}$ is a *pedigree on* X. The set V_0 of vertices of $\mathscr{P}$ of indegree 0 is sometimes referred to as the *founders* of the pedigree. Note that for any *finite pedigree* (i.e. any pedigree for which $|V|$ is finite), the set V_0 is necessarily non-empty. An example of a pedigree on $X = \{1, 2, 3, 4\}$ is shown in Fig. 1, in which the round vertices denote females, square vertices denote males, V_0 consists of the 12 individuals at the top, and all arcs are oriented downwards.

To study pedigrees it is sometimes helpful to define a more general type of acyclic digraph, which we will call a (*general*) *pedigree*. This has condition (P) replaced by the weaker condition:

(P*) if v has positive indegree then either v has exactly two incoming arcs, say (u, v) and (u', v), where $u \in M$ and $u' \in F$, or v has one incoming arc.

Henceforth, 'pedigree' will mean a general pedigree, unless we specifically use the prefix 'strict'. Two other definitions that are widely used in this paper are:

- For an arc (u, v) of any pedigree we say that v is a *child* of u and u is a *parent* of v.
- For any vertex v of $\mathscr{P}$ let $\gamma(v) \in \{M, F\}$ indicate the gender class that v belongs to.

It is important to formalize what it means for two pedigrees to 'be essentially the same' (isomorphic) where our set of extant individuals X is known (and so effectively labelled) but where we do not care about the labelling of the ancestral individuals. We may wish to either take account of, or ignore, the sex of ancestral individuals in a pedigree, and so we introduce the following two definitions. Two pedigrees on X, $\mathscr{P}_1 = (V_1, A_1)$ and $\mathscr{P}_2 = (V_2, A_2)$, are *isomorphic*, written $\mathscr{P}_1 \cong \mathscr{P}_2$ if there is a bijective map $\phi : V_1 \to V_2$ for which $(v_1, v_2) \in A_1 \Leftrightarrow (\phi(v_1), \phi(v_2)) \in A_2$ (i.e. ϕ is a digraph isomorphism) and ϕ is the identity map when restricted to X. If in addition $\phi(M_1) = M_2, \phi(F_1) = F_2$ (where M_i and F_i is the gender-based bi-partition of V_i) then we say that $\mathscr{P}_1$ and $\mathscr{P}_2$ are *gender isomorphic*, written $\mathscr{P}_1 \cong_g \mathscr{P}_2$.

For example, consider the the two pedigree graphs in Fig. 2 which show two individuals $\{x, y\}$ that share the same four grandparents. These two pedigrees are non-isomorphic (and therefore not gender isomorphic). What makes this example interesting is that these two pedigrees are similar in many other respects—for example the distribution of path lengths from x to y is the same, and the ancestry of each individual (x or y) is the same for both pedigrees.

Pedigree graphs have a number of attractive graph-theoretic properties—for example, it is easily seen that they can be properly 3-coloured (Semple and Steel, 2003). Furthermore, any graph obtained from a pedigree by adding edges between only M and F vertices can be

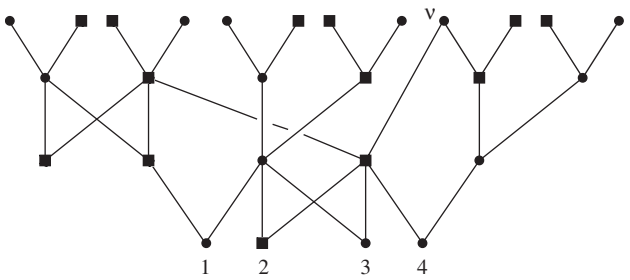

Fig. 1. A pedigree on $X = \{1, 2, 3, 4\}$.

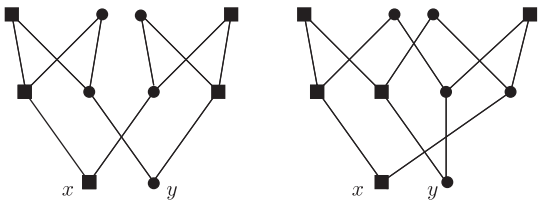

Fig. 2. Two pedigree graphs that are non-isomorphic even when gender is ignored.

properly 4-coloured (Thomas, 1993). This last fact follows from the observation that the induced subgraph of any pedigree $\mathscr{P}$ on M is a forest (similarly the induced subgraph of $\mathscr{P}$ on F) and any tree (hence forest) can be properly 2-coloured. This 4-colourability result implies that certain pedigree graphs have small 'tree-width', which can be useful for various statistical calculations. In general, however a pedigree may have large tree-width, and so, not surprisingly, the calculation of certain statistical properties of pedigrees can be computationally complex (see e.g. Aceto et al., 2004; Piccolboni and Gusfield, 2003; Thomas, 1993).

1.2. Further definitions (restricted pedigree, depth)

We will be interested in restricting pedigrees to subsets of vertices. To make this more precise, given a pedigree $\mathscr{P} = (V, A)$, and a subset U of V, let $\mathscr{P}|U$ denote the induced digraph which has vertex set U and arc set $\{(v, v') \in A : v, v' \in U\}$. We call $\mathscr{P}|U$ the *restriction of* $\mathscr{P}$ *to* U. For a subset S of X and a pedigree $\mathscr{P}$ on X, with vertex set V, let $V(S)$ denote the set of vertices v of $\mathscr{P}$ for which there is a path from v to at least one element of S, and let $\mathscr{P}(S)$ denote the restriction of $\mathscr{P}$ to $V(S)$; that is,

$$\mathscr{P}(S) = \mathscr{P}|V(S).$$

For brevity we will often write $\mathscr{P}(x_1, \ldots, x_r)$ in place of $\mathscr{P}(\{x_1, \ldots, x_r\})$.

For a vertex v in a pedigree, define the *depth* of v, denoted $\Delta(v)$, to be the length of the longest directed path (number of arcs) from v to an element of X. For example, in Fig. 1, the vertex v indicated has depth 3.

The *depth* of $\mathscr{P}$, denoted $\Delta(\mathscr{P})$ is the largest depth of any vertex in $\mathscr{P}$. A pedigree $\mathscr{P}$ has *discrete-step generations* if the depth of any vertex also equals the length of the shortest path (and therefore the length of every path) from that vertex to any element of X (as in Fig. 2, but not in Fig. 1); any such pedigree can be 2-coloured (i.e. is bipartite if regarded as an undirected graph), since we can assign to each vertex its depth modulo 2 to obtain a proper colouring. In this paper we generally do not restrict ourselves to discrete-step generation pedigrees (many of the proofs simplify considerably, however the discrete-step generation assumption is very strong).

Given a pedigree $\mathscr{P}$ on X, and a subset S of X let $\mathscr{P}(S; \Delta \leqslant d)$ denote the restriction of $\mathscr{P}(S)$ to the set of those vertices v of depth at most d in $\mathscr{P}(S)$.

To illustrate these ideas for the pedigree $\mathscr{P}$ on $X = \{1, 2, 3, 4\}$ shown in Fig. 1, the pedigrees $\mathscr{P}(\{2, 3\}; \Delta \leqslant 2)$ and $\mathscr{P}(X; \Delta \leqslant 2)$ are displayed in Fig. 3(a), (b), respectively. Notice that the vertex labelled v in Fig. 1 appears in $\mathscr{P}(\{2, 3\}; \Delta \leqslant 2)$ (as shown) but not in $\mathscr{P}(X; \Delta \leqslant 2)$.

Note also that if $\mathscr{P}$ is a general (or strict) pedigree on X, and $S \subseteq X$ then $\mathscr{P}(S; \Delta \leqslant d)$ is a general pedigree on S; however if $\mathscr{P}$ is a strict pedigree on X, then $\mathscr{P}(S; \Delta \leqslant d)$ may fail to be a strict pedigree on X. This is because there may exist a vertex at depth $d - 1$ which has precisely one parent

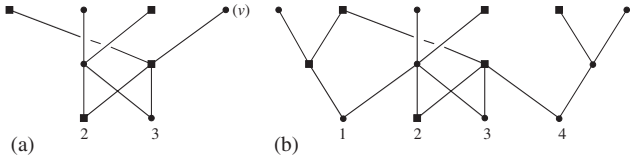

Fig. 3. The restricted pedigrees: (a) $\mathscr{P}(\{2, 3\}; \Delta \leqslant 2)$ and (b) $\mathscr{P}(X; \Delta \leqslant 2)$ for the pedigree $\mathscr{P}$ shown in Fig. 1 (see text for details).

at depth d (this would be impossible if $\mathscr{P}$ had discrete-step generations, but could occur otherwise, as Fig. 3(b) illustrates).

2. Pedigree reconstruction by pairwise amalgamation

Our first result shows how any pedigree, up to any depth, can be reconstructed from pedigrees of that depth on pairs of individuals from X. Note that in this result individuals in the pedigree that have no surviving descendants in X will not be part of any reconstructed pedigree.

Theorem 2.1. *Let $\mathscr{P}$ be a (general) finite pedigree on X. Then,*

(i) *$\mathscr{P}(X)$ can be reconstructed (up to gender isomorphism) from the collection of pedigrees $\mathscr{P}(x, x')$ of all pairs $x, x' \in X$ (described up to gender isomorphism).*
(ii) *For any non-negative integer d the (general) pedigree $\mathscr{P}(X; \Delta \leqslant d)$ can be reconstructed (up to gender isomorphism) from the collection of pedigrees $\mathscr{P}(\{x, x'\}; \Delta \leqslant d)$ of all pairs $x, x' \in X$ (described up to gender isomorphism).*

In either case the reconstruction can be achieved by an algorithm that runs in polynomial time (in the number of vertices of $\mathscr{P}$ and of $\mathscr{P}(X; \Delta \leqslant d)$ respectively).

Proof. Part (i) follows immediately from part (ii) by taking $d = \Delta(\mathscr{P})$ so it suffices to establish part (ii). We may assume that $X = \{1, 2, \ldots, n\}$. We will establish the following:

Claim 1. *For $m \in \{2, \ldots, n\}$, $\mathscr{P}(\{1, \ldots, m\}; \Delta \leqslant d)$ can be determined up to gender isomorphism from the gender isomorphism classes of $\mathscr{P}(1, \ldots, m-1; \Delta \leqslant d)$ and the collections $\mathscr{P}(\{i, m\}; \Delta \leqslant d)$ for choices of $i \in \{1, \ldots, m-1\}$.*

For this task we use induction on m. First note that for $m = 2$ Claim 1 holds trivially. To establish the induction step we introduce some further notation. Suppose $\mathscr{P}$ is a pedigree on a set $S \subseteq X$. For any vertex v of $\mathscr{P}$ recall that $\gamma(v) \in \{M, F\}$ indicates the gender class that v belongs to. Given $s \in S$, a *gender path* from s to a vertex w of $\mathscr{P}$ is any sequence $(\gamma(v_1), \ldots, \gamma(v_k))$ for $k \geqslant 1$, where (v_1, v_0), $(v_2, v_1), \ldots, (v_k, v_{k-1})$ are arcs of $\mathscr{P}$ and $v_0 = s, v_k = w$. Note that, for any sequence $g = (g_1, \ldots, g_k) \in \{M, F\}^k$ and any $s \in S$, there is at most one vertex z of $\mathscr{P}$ for which g is the gender sequence from s to z; we call such a vertex z the

vertex specified by taking the gender path g from s and denote it by $z(s,g)$. For example, in Fig. 1, the vertex v shown can be represented as $z(4,(M,F))$, $z(4,(F,M,F))$, $z(3,(M,F))$ or $z(2,(M,F))$.

Returning to the induction step for Claim 1, let $\mathscr{P}_1 = (V_1, A_1)$ be gender isomorphic to $\mathscr{P}(\{1,\ldots,m-1\}; \Delta \leqslant d)$ and $\mathscr{P}_2 = (V_2, A_2)$ be gender isomorphic to $\mathscr{P}(\{m\}; \Delta \leqslant d)$. To simplify the following construction we will regard V_1 and V_2 as disjoint sets (this is possible because we are dealing with isomorphism classes, and it is convenient for what is to follow).

Next we describe a relation $\sim$ on $V_1 \times V_2$ (the point of this relation will be to identify two vertices if they represent the same individual in $\mathscr{P}(1,\ldots,m)$).

For $v_1 \in V_1$ and $v_2 \in V_2$ we determine whether the following condition holds for any $i \in \{1,\ldots,m-1\}$ that is a descendant of v_1: Select a gender path g from i to v_1 in $\mathscr{P}_1$ and a gender path g' from m to v_2 in $\mathscr{P}_2$. Now, in $\mathscr{P}(\{i,m\}; \Delta \leqslant d)$ let $z(i,g)$ denote the vertex specified by taking gender path g from i, and let $z(m,g')$ denote the vertex specified by taking the gender path g' from m. Set $v_1 \sim v_2$ if and only if $z(i,g) = z(m,g')$. This relation $\sim$ is well defined (i.e. independent of the choice of i, g and g').

We now construct a digraph $\mathscr{D}$ with vertex set V obtained from $V_1 \cup V_2$ by identifying $v_1 \in V_1$ and $v_2 \in V_2$ whenever $v_1 \sim v_2$. Since each vertex $v \in V$ corresponds to a vertex in V_1 or V_2 or in both, we can define the arc set A of $\mathscr{D}$ to be those pairs $(v,v') \in V \times V$ for which correspond to a pair of vertices in V_1 (or in V_2) that form an arc in A_1 (or in A_2, respectively). Then $\mathscr{D} = (V, A)$ is a general pedigree on $\{1,\ldots,m\}$, and so we may further let (V', A') denote the restriction of this pedigree to vertices of depth at most d.

Consider the following bijective map from the vertices of $\mathscr{P}(\{1,\ldots,m\}; \Delta \leqslant d)$ to V'. Each vertex v in $\mathscr{P}(\{1,\ldots,m\}; \Delta \leqslant d)$ either has a descendant in $\{1,\ldots,m-1\}$, in which case v corresponds to a unique vertex in $V_1 \cap V'$, or v has m as a descendant and in addition v has no descendant in $\{1,\ldots,m-1\}$, in which case v corresponds to a unique vertex in $V_2 \cap V'$ that is not $\sim$ related to any vertex in V_1. This map provides a gender isomorphism from $\mathscr{P}(\{1,\ldots,m\}; \Delta \leqslant d)$ to (V', A'), and thereby establishes Claim 1.

Setting $m = n$ in Claim 1, $\mathscr{P}(\{1,\ldots,m\}; \Delta \leqslant d)(= \mathscr{P}(X; \Delta \leqslant d))$ can be reconstructed up to gender isomorphism from the collection of pedigrees $\mathscr{P}(\{x,x'\}; \Delta \leqslant d)$ for $x, x' \in X$. Furthermore, the construction described above can clearly be implemented in polynomial time. This completes the proof. □

2.1. Remark

Theorem 2.1 is no longer true if one replaces the phrase 'gender isomorphism' by 'isomorphism' throughout. To see this consider Fig. 4. We can construct two pedigrees $\mathscr{P}, \mathscr{Q}$ on $X := \{x,y,z\}$ from this graph as follows. To form $\mathscr{P}$ identify vertex p and p'. To form $\mathscr{Q}$ identify vertex p and p''. Note that $\mathscr{P}(X)$ and $\mathscr{Q}(X)$ are not isomorphic, however

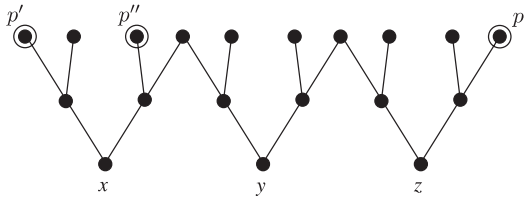

Fig. 4. Example to illustrate that Theorem 2.1 requires gender isomorphism.

$\mathscr{P}(x_1,x_2)$ is isomorphic to $\mathscr{Q}(x_1,x_2)$ for all $x_1, x_2 \in X$. Note also that, in this example, a consistent assignment of sexes is possible in both cases, but since we are dealing with isomorphism and not gender isomorphism the details are not relevant.

3. Reconstructing pedigrees from 'link' and 'lasso' data

Theorem 2.1 requires as input entire pedigrees for pairs of extant individuals. In this section, we show that a pedigree can be reconstructed (up to gender isomorphism) from more atomic information—namely from gender information about certain circuits in the pedigree. To describe this we begin with some definitions.

Let $\mathscr{P} = (V, A)$ be a pedigree on X. Given vertices $(u,u') \in V \times V$ a *pedigree link for* (u,u'), is a pair (W, W') where $W = (w_0, w_1, \ldots, w_j)$, $W' = (w'_0, w'_1, \ldots, w'_k)$ are sequences of vertices, with $j,k \geqslant 0$ and for which the following two conditions apply:

- $w_j, w_{j-1}, \ldots, w_0$ is a directed path from w_j to $w_0 = u$, and $w'_k, w'_{k-1}, \ldots, w'_0$ is a directed path from w'_k to $w'_0 = u'$, and in addition $w_j = w'_k$.
- These two directed paths share no arcs, and the only vertices they share are $w_j = w'_k$, and, in the special case where $u = u'$, $w_0 = w'_0$.

Informally, a pedigree link for (u,u') can be viewed as two separate lines of ancestry that trace back in time from v and w to some common ancestor. A generic pedigree link is shown schematically in Fig. 5(a). Note the definition allows for a pedigree link for the pair (u,u), including the trivial link $((u),(u))$. The *depth* of a pedigree link $((w_0, w_1, \ldots, w_j), (w'_0, w'_1, \ldots, w'_k))$ is $\max\{j,k\}$.

A *pedigree lasso* on a vertex u is an ordered triple (U, W, W') where

- $U = (u_0, u_1, \ldots, u_k)$, with $k \geqslant 1, u_0 = u$ and $(u_{i+1}, u_i) \in A$ for each $i \in \{0,\ldots,k-1\}$;
- (W, W') is a pedigree link for (u_k, u_k).

The *depth* of this pedigree lasso is $k + \text{depth}(W, W')$, and in case $\text{depth}(W, W') = 0$ we say the lasso is *trivial*. Informally, a pedigree lasso on u can be viewed as a line of ancestry that travels back in time from u to some ancestor, and that then either ends or splits into two separate lines of ancestry that eventually meet again at a

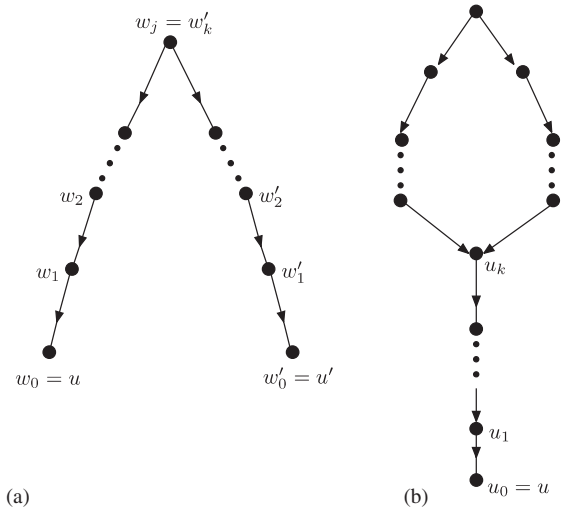

Fig. 5. Generic pedigree link (a), and pedigree lasso (b).

more distant ancestor. A generic pedigree lasso is shown schematically in Fig. 5(b).

3.1. Gender sequences

We will be interested just in the gender sequences associated with pedigree links and lassos, and the following definitions make this precise.

Let Γ denote the collection of finite sequences over $\{M, F\}$, and for $\gamma \in \Gamma$, let $|\gamma|$ denote the length of γ. Given any sequence $Z = (z_0, z_1, \ldots, z_r)$ of vertices of $\mathscr{P}$ let $\gamma(Z) \in \Gamma$ denote the 'pruned' gender sequence $(\gamma(z_1), \ldots, \gamma(z_r))$ (we say 'pruned' because we omit $\gamma(z_0)$; in case $r = 0$ we will simply write $\gamma(Z) = -$).

For an ordered pair (x, y) of elements of X, let $Lk(x, y)$ denote the set of pairs $(\gamma^1, \gamma^2) \in \Gamma \times \Gamma$ of (pruned) gender sequences $(\gamma(W), \gamma(W'))$ over all pedigree links (W, W') for (x, y). Note that $(\gamma^1, \gamma^2) \in Lk(x, y)$ if and only if $(\gamma^2, \gamma^1) \in Lk(y, x)$.

To illustrate this concept consider the two pedigrees shown in Fig. 2. The pedigree shown on the left has:

$$Lk(x, y) = \{((M, M), (F, M)), ((M, F), (F, F)), ((F, F), (M, F)), ((F, M), (M, M))\},$$

while the pedigree shown on the right has:

$$Lk(x, y) = \{((M, M), (M, M)), ((M, F), (F, F)), ((F, F), (M, F)), ((F, M), (F, M))\}.$$

Similarly, for an element x of X, let $L(x)$ denote the set of triples $(\gamma^1, \gamma^2, \gamma^3) \in \Gamma^3$ of (pruned) gender sequences $(\gamma(U), \gamma(W), \gamma(W'))$ over all pedigree lassos (U, W, W') on x. For example, for the pedigree graphs in Fig. 2, $L(x)$ and $L(y)$ consist just of (trivial) triples such as $((F), -, -)$ and $((M, F), -, -)$.

Given a pedigree $\mathscr{P}$ on X we refer to an element (γ^1, γ^2) of $Lk(x, y)$ as a *gender link* for (x, y) (of depth $\max\{|\gamma^1|, |\gamma^2|\}$) and the associated function $Lk : X \times X \to 2^{\Gamma}$ as the *link function* of the pedigree. Similarly we refer to an element $(\gamma^1, \gamma^2, \gamma^3)$ of $L(x)$ as a *gender lasso* on x (of depth $|\gamma^1| + \max\{|\gamma^2|, |\gamma^3|\}$) and the associated function: $L : X \times X \to 2^{\Gamma}$ as the *lasso function* of the pedigree.

Note that for pedigrees with discrete-step generations the description of a link and lasso can be simplified. Namely, one does not need to specify separately the gender type of the two arc-disjoint paths from the ancestral vertex, and can instead use the sequence of genders encountered in a (necessarily even length) cycle from one element of X to another.

Theorem 3.1. *Let $\mathscr{P}$ be any (general) finite pedigree, $\mathscr{P}$, on X. Then $\mathscr{P}(X)$ is determined up to gender isomorphism by the pair (L, Lk) (the link and lasso functions of $\mathscr{P}$).*

Proof. First observe that the link and lasso function of $\mathscr{P}$ is exactly the same as the link and lasso function of $\mathscr{P}(X)$ and so we may assume for the remainder of this proof that $\mathscr{P} = \mathscr{P}(X)$.

We use induction on the number $k = k(\mathscr{P})$ of arcs $\mathscr{P}$. The idea for the inductive step will be to remove one carefully-chosen arc from $\mathscr{P}$, and show how the link and lasso functions on the resulting pedigree are determined by the original link and lasso functions (working over general, rather than strict pedigrees is useful for the proof, since removing an arc may create an individual with just one parent).

First note that for $k = 0$ we have $\mathscr{P} = \mathscr{P}(X) = (X, \emptyset)$ and clearly we can determine that $k = 0$ from L. Thus in case $k = 0$ we can reconstruct $\mathscr{P}(X)$, from L. For the induction step, suppose that Theorem 3.1 holds for all $\mathscr{P}$ with $k(\mathscr{P})$ less than $K \geqslant 1$ and that $\mathscr{P}$ is a finite pedigree on X with $k(\mathscr{P}) = K$.

Note that a parent of an element of X may not necessarily have depth 1 (unless one assumes discrete-step generations). Nevertheless we claim that when $k(\mathscr{P}) > 0$ there always exists a vertex v of $\mathscr{P}$ that has depth precisely 1—indeed we may take v to be any vertex in $\mathscr{P}$ that does not lie in X, and that has minimal depth d. Clearly $d \geqslant 1$ and to see that we must have $d = 1$, suppose that $d > 1$. Select a shortest path from v to an element of X, and let v' be the penultimate vertex of $\mathscr{P}$ in this path. By assumption, the depth of v' is at least d, and v' is different to v. However this implies that the depth of v is at least $d + 1$, a contradiction. It follows that there is indeed a vertex v of $\mathscr{P}$ of depth 1; that is, an individual all of whose children either lie in X or do not leave any descendants in X.

Note that the elements $x \in X$ that have a parent of depth 1 are precisely the elements $x \in X$ that satisfy the following condition for some $\gamma_x \in \{M, F\}$:

For all $y \in X, \gamma \in \Gamma$, if $((\gamma_x), \gamma) \in Lk(x, y)$ then $\gamma = (\gamma_x)$.

Thus by using just Lk we can select an element x satisfying this last condition (for a particular γ_x)—as we have shown, such an x must exist. In addition, we say that x is γ_x-*solo* if x is the only child of the γ_x-parent of x, and we say that x is a *miracle* if x has only one parent (i.e. the γ_x-parent). Now,

consider the general pedigree $\mathscr{P}'$ obtained from $\mathscr{P}$ by deleting the arc from the γ_x-parent of x to x, and in addition (i) if x is a miracle, delete x; and (ii) if x is γ_x-solo then assign a label (not present in X) to the γ_x-parent, for example $z = z(x, \gamma_x)$. Let X' denote the set of vertices of $\mathscr{P}'$ of out-degree 0. Then $X - \{x\} \subseteq X' \subseteq X \cup \{z\}$. Moreover, X' contains x iff x is not a miracle, and X' contains z iff x is γ_x-solo. We will regard $\mathscr{P}'$ as a pedigree *on* the labelled set X'.

We claim that the lasso and link functions for $\mathscr{P}'$—call them L' and Lk', respectively—are determined by L and Lk. To see this observe that for $u, w \in X - \{x\}$ we have

$$Lk'(u, w) = Lk(u, w),$$

$$L'(u) = L(u).$$

Also if $z \in X'$ (i.e. x is γ_x-solo) then

$$(\gamma^1, \gamma^2) \in Lk'(z, z) \Leftrightarrow ((\gamma_x), \gamma^1, \gamma^2) \in L(x); \quad (1)$$

and

$$(\gamma^1, \gamma^2, \gamma^3) \in L'(z) \Leftrightarrow (\gamma_x \cdot \gamma^1, \gamma^2, \gamma^3) \in L(x),$$

where $\cdot$ denotes concatenation; while for $w \in X - \{x\}$ we have

$$(\gamma^1, \gamma^2) \in Lk'(z, w) \Leftrightarrow (\gamma_x \cdot \gamma^1, \gamma^2) \in Lk(x, w).$$

Now suppose that $x \in X'$ (i.e. x is not a miracle). Then,

$$(\gamma^1, \gamma^2, \gamma^3) \in L'(x) \Leftrightarrow (\gamma^1, \gamma^2, \gamma^3) \in L(x) \text{ and } \gamma^1_1 \neq \gamma_x$$

and for $y \in X' - \{z\}$,

$$(\gamma^1, \gamma^2) \in L'(x, y) \Leftrightarrow (\gamma^1, \gamma^2) \in L(x, y) \text{ and } \gamma^1_1 \neq \gamma_x$$

(where γ^1_1 refers to the first term of γ^1). Finally, if both $x, z \in X'$ then

$$(\gamma^1, \gamma^2) \in L'(x, z) \Leftrightarrow (\gamma^1, \gamma_x \cdot \gamma^2) \in L(x, x).$$

Thus, since $k(\mathscr{P}') = k(\mathscr{P}) - 1$, it follows by the inductive hypothesis that $\mathscr{P}'(X')$ is determined up to gender isomorphism by L', Lk' and so by L, Lk. Now, the pedigree obtained from $\mathscr{P}'(X')$ by (i) re-introducing x if x is not an element of X', and (ii) adding an arc from z to x, is (gender isomorphic to) $\mathscr{P}(X)$. This establishes the induction step, and thereby the proof of Theorem 3.1. □

Remarks.

- Although Theorem 3.1 is related to Theorem 2.1(i) neither result follows immediately from the other, and the type of induction used in both cases is quite different. In particular, Theorem 3.1 makes no claim concerning a polynomial-time algorithm for reconstructing $\mathscr{P}$ since $L(x)$ can clearly be of exponential size in the number of vertices of $\mathscr{P}$. Furthermore, the analogue of Theorem 2.1(ii) (i.e. depth-restriction) does not hold in the setting of Theorem 3.1; for example the two pedigrees $\mathscr{P}$ and $\mathscr{P}'$ in Fig. 6 have the same link and lasso functions for links and lassos of depth at most 1, yet $\mathscr{P}(X; \Delta \leqslant 1)$ is not gender isomorphic to $\mathscr{P}'(X; \Delta \leqslant 1)$.

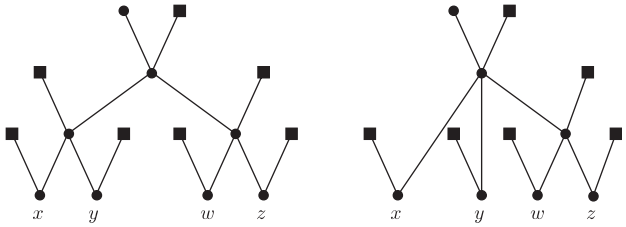

Fig. 6. Two pedigrees with identical link and lasso functions of depth at most 1, yet with non-isomorphic restriction to vertices of depth at most 1.

- It is clear that both L and Lk are required in order to reconstruct $\mathscr{P}$; nevertheless it is interesting to ask how much can be determined about a pedigree from just its link function. Specifically, for a general pedigree $\mathscr{P}$ on X, let VL denote the set of all vertices of $\mathscr{P}$ that lie on at least one pedigree link for a pair of elements of X (including pairs for the form (x, x)). It might be conjectured that $\mathscr{P}|VL$ can be reconstructed just from the Lk, however a counterexample can be constructed to show that in general this is not the case (the problem with extending the inductive argument employed for Theorem 3.1 is that Eq. (1) requires the lasso function L for determining a value of Lk'). How much that Lk determines concerning $\mathscr{P}|VL$ remains an interesting question.

4. Pedigree enumeration and limits to pedigree reconstruction from segregating sequence sites for X

In this section, we derive a lower bound on the number of isomorphism classes of pedigrees on X that have the property that each individual in the pedigree has depth at most d. We then describe some consequences for pedigree reconstruction. We show that the number of isomorphism classes of such pedigrees on X grows at least as fast as the function $n^{\beta nd}$ where $n = |X|$ and β is a positive constant, and that this bound applies even if one imposes further restrictions (e.g. a bound on the number of children each individual can have). To obtain this lower bound we will see that we need only consider a very particular sub-class of pedigrees over X. Note that in this section we use isomorphism in the more general sense; that is, we do not distinguish between sexes.

Theorem 4.1. *The number of isomorphism classes of pedigrees on X of depth d grows at least at the rate $n^{\beta nd}$ for a positive constant β, and $n = |X|$. Furthermore, this bound holds even if we further insist that any combination of the following hold*:

- *Each individual vertex in the pedigree has at least one descendant in X,*
- *The number of vertices at any given depth is at most (a constant ($\geqslant 1$) times) $|X|$,*
- *The pedigree has discrete-step generations,*
- *The number of children each individual has is at most* 3.

To prove this result we first note that since we are concerned with lower bounds it suffices to count any subclass of pedigrees of depth d satisfying the four extra conditions specified in the theorem, and our choice is dictated largely by computational convenience. Thus, consider discrete generation pedigrees of depth d and constant population size $n = 3k$ and which satisfy the following requirement:

(3X) In each generation the population can be partitioned into k sets of size 3, so that the following conditions hold: within each triple, two individuals share the same parents—say p, p'—while the third individual has just one of these as a parent—say p—as well as a second parent, say p''.

An example for $k = d = 2$ is shown in Fig. 7 for $X = \{1, 2, 3, 4, 5, 6\}$ partitioned into the triples $\{1, 2, 3\}$ and $\{4, 5, 6\}$. The three circled vertices at depth one form a triple in the partition at that level. Condition (3X) may appear strange, however it suffices to provide a computable lower bound on the number of isomorphism classes of pedigrees on X.

Theorem 4.1 follows as an immediate consequence of the following result (noting that pedigrees that satisfy (3X) satisfy all four extra conditions mentioned in Theorem 4.1).

Proposition 4.2. *Let $N(d, k)$ denote the number of isomorphism classes of discrete generation pedigrees of depth d and constant population size $n = 3k$ and that satisfy condition* (3X). *Then*

$$|N(d,k)| = \left(\frac{(3k)!}{2^k k!}\right)^d.$$

Furthermore, $|N(d,k)| \geqslant (k^2\beta)^{kd}$ where $\beta = 27e^{-2}/2 \cong 1.83$.

Proof. We use induction on d to establish the result.

For $d = 1$ first note that the number of ways of partitioning the set X (of size $3k$) into k sets of size 3 is $\frac{(3k)!}{(3!)^k k!}$ (Ross and Wright, 1999, p. 305). For each set of size 3 we must select one individual as the candidate in the triple for the third individual mentioned in condition (3X) (with parents p, p''). The selection of these triples and such an individual for each triple specifies the pedigree up to isomorphism, and so

$$N(1,k) = \frac{(3k)!}{(3!)^k k!} \times 3^k = \left(\frac{(3k)!}{2^k k!}\right)^1,$$

which establishes the result for $d = 1$.

Now suppose $d > 1$ and let $\mathscr{P}$ be a pedigree on X that satisfies (3X) and has depth d. Note that a labelling of the individuals at depth $d' - 1 \in \{0, \ldots, d-1\}$ specifies a labelling of individuals at depth d' which is dependent only on the isomorphism class. This labelling associates to each individual at depth d' the set of labels of children it has at height $d' - 1$ (beginning with the labelled set X at $d' = 0$). For example, the third, fifth and sixth vertices (from left to right) in the top layer of Fig. 7 receive the labels,

$$\{\{2,3\},\{4,5\}\}; \{\{2,3\},\{4,5\},\{6\}\}; \text{ and } \{\{6\}\}.$$

Since the labels at level $d - 1$ are all distinct, and this labelling depends only on (and also determines) the isomorphism class of $\mathscr{P}$ it follows that

$$N(d,k) = N(d-1,k) \cdot N(1,k),$$

which establishes the induction step.

The inequality in the second part of Theorem 4.2 follows from the equality by using a form of Stirling's formula that states that $m! = c(m)m^m e^{-m}\sqrt{2\pi m}$ where $1 < c(m) < \exp(1/12m)$ (Robbins, 1955) applied to $m = k, 3k$. □

Note that the bound in Theorem 4.1 (or Proposition 4.2) would also apply to counting gender isomorphism classes, but here we are using the weaker notion of isomorphism (defined earlier) which does not distinguish sexes.

Proposition 4.2 has the following implication for the reconstruction of pedigrees from segregating sites. Suppose we have s sites in DNA that show variation across a population (such sites are called 'segregating sites'). Suppose that we have some model describing how the structure of a pedigree influences the segregating sites (for example a model that combines mutation and recombination). Then we can ask how many segregating sites we need in order to reconstruct a pedigree back d generations. This will in general depend very much on the details of the model, however we can set generic lower bounds (independent of the details of the model) by simply comparing the number of possible data sets involving s segregating sites with the number of pedigrees of depth d for a population of size n. An upper bound on the former number is 4^{ns} (the value 4 since DNA has four bases) while a lower bound on the latter number is given by the second part of Proposition 4.2. Applying then an information-theoretic argument (Steel and Székely, 1999, Theorem 2.1(ii)) we obtain the following.

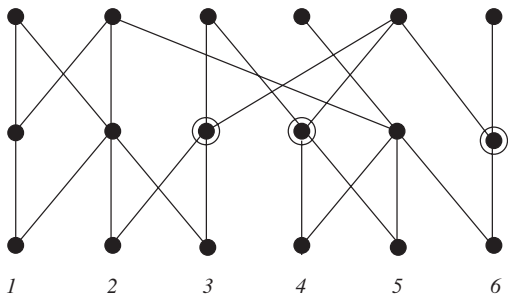

Fig. 7. Condition (3X) for $k = d = 2$ and $X = \{1, 2, 3, 4, 5, 6\}$ (see text for details).

Corollary 4.3. *Suppose s is the number of segregating DNA sequence sites that are required (by any method and any model of site evolution) to reconstruct correctly with probability larger than* 0.5 *each pedigree of depth d and*

constant population size n. Then,

$$s \geqslant \frac{d}{3}\log_2\left(\frac{n}{3}\right).$$

For example, to reconstruct the pedigree of a human population of constant size 3 million for 300 generations (approx. 7000 years) would require an absolute minimum of 2000 segregating sites. In reality the number is likely to be much higher due to the under-counting of pedigree isomorphism classes by imposing condition (3X), and due to stochastic aspects of any model of site evolution.

5. Concluding comments

The results described above raise many questions; both combinatorial and stochastic.

One combinatorial question (concerning compatibility) is to determine for a set of pedigrees on $Y_1, \ldots, Y_k$ whether there exists a pedigree $\mathscr{P}$ on $X = \bigcup_{i=1}^{k} Y_i$ that contains each of the input pedigrees as an isomorphic copy.

Remark (2.1) suggests another combinatorial question. Does there exists some value $r > 2$ for which, for any two pedigrees $\mathscr{P}, \mathscr{Q}$ on X,

$$\mathscr{P}(X) \cong \mathscr{Q}(X) \Leftrightarrow \mathscr{P}(S) \cong \mathscr{Q}(S) \text{ for all } S \subseteq X, |S| \leqslant r?$$

An enumeration question would be to find better lower (or upper) bounds for the number of isomorphism classes of pedigrees on X; or perhaps even exact formulae. For this, the ideas developed in (Thomas and Cannings, 2003, 2004) may be useful.

In practice, pedigree reconstruction is likely to be based on sequences that have evolved under models of site evolution (involving recombination and mutation). The bound we have described in terms of the number of segregating sites is likely to be a considerable underestimate for the number of segregating sites that may be required for accurate pedigree reconstruction, however we can ask whether the same *rate of growth* of s, namely $d\log(n)$, can be achieved. There is a curious parallel in the field of phylogeny reconstruction where equally primitive counting arguments showed that sequences must grow at the rate (at least) $\log(n)$ in order to accurately reconstruct trees with n leaves. Yet it turned out that under certain models this lower bound was indeed the actual rate of growth required, at least for most trees (Erdös et al., 1999). Whether such a fortuitous outcome is the case for pedigrees is unlikely, though a reasonable topic for further investigation.

Acknowledgements

We thank Rune Lyngsoe for many helpful comments on an earlier version of this manuscript. We also thank the New Zealand Marsden Fund (UOC310) and the Allan Wilson Centre for supporting this research.

References

Aceto, L., Hansen, J., Ingólfsdöttir, A., Johnsen, J., Knudsen, J., 2004. The complexity of checking consistency of pedigree information and related problems. J. Comput. Sci. Technol. 19 (1), 42–59.

Bininda-Emonds, O.R.P., 2004. Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. Kluwer Academic Publishers, Dordrecht.

Cannings, C., Thompson, E.A., 1981. Genealogical and Genetic Structure. Cambridge University Press, Cambridge, UK.

Erdös, P.L., Székely, L.A., Steel, M., Warnow, T., 1999. A few logs suffice to build (almost) all trees (I). Random. Struct. Algorithms 14, 153–184.

Mossel, E., Steel, M., 2005. How much can evolved characters tell us about the tree that generated them? In: Gascuel, O. (Ed.), Mathematics of Evolution and Phylogeny. Oxford University Press, Oxford, pp. 384–412.

Piccolboni, A., Gusfield, D., 2003. On the computational complexity of fundamental computational problems in pedigree analysis. J. Comput. Biol. 10 (5), 763–773.

Robbins, H., 1955. A remark on Stirling's formula. Amer. Math. Monthly 62, 26–29.

Ross, K.A., Wright, C.R.B., 1999. Discrete Mathematics, fourth ed. Prentice-Hall, Englewood Cliffs, NJ.

Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford.

Sober, E., Steel, M., 2002. Testing the hypothesis of common ancestry. J. Theor. Biol. 218, 395–408.

Steel, M.A., Székely, L.A., 1999. Inverting random functions. Ann. Combin. 3, 103–113.

Thomas, A., 1993. A note on the four-colourability of pedigrees and its consequences for probability calculations. Stat. Comput. 3, 51–54.

Thomas, A., Cannings, C., 2003. Enumeration and simulation of marriage node graphs on zero-loop pedigrees. Math. Med. Biol. 20, 261–275.

Thomas, A., Cannings, C., 2004. Simulating realistic zero loop pedigrees using a bipartite Prüfer code and graphical modelling. Math. Med. Biol. 21, 335–345.

Thompson, E.A., 2000. Statistical inference from genetic data on pedigrees. Institute of Mathematical Statistics.