Journal of Theoretical Biology 412 (2017) 8–16

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Reconstructing pedigrees using probabilistic analysis of ISSR amplification☆

Loïc Chaumont[a,*], Valéry Malécot[b], Richard Pymar[c], Chaker Sbai[d]

[a] LAREMA – UMR CNRS 6093, Université d'Angers, 2 bd Lavoisier, 49045 Angers Cedex 01, France
[b] IRHS – UMR 1345, Agrocampus Ouest Angers, 2 rue Le Nôtre, 49045 Angers Cedex 01, France
[c] Department of Statistical Science – University College London, Gower Street, London WC1E 6BT, United Kingdom
[d] PEGASE – UMR 1348, Agrocampus Ouest Rennes, 65 rue de Saint-Brieuc, CS 84215, 35042 Rennes Cedex, France

ARTICLE INFO

MSC:
92D25
92D10
60F15

Keywords:
Pedigree
ISSR amplification
Law of reproduction
Gene frequency

ABSTRACT

Data obtained from ISSR amplification may readily be extracted but only allows us to know, for each gene, if a specific allele is present or not. From this partial information we provide a probabilistic method to reconstruct the pedigree corresponding to some families of diploid cultivars. This method consists in determining for each individual what is the most likely couple of parent pair amongst all older individuals, according to some probability measure. The construction of this measure bears on the fact that the probability to observe the specific alleles in the child, given the status of the parents does not depend on the generation and is the same for each gene. This assumption is then justified from a convergence result of gene frequencies which is proved here. Our reconstruction method is applied to a family of 85 living accessions representing the common broom *Cytisus scoparius*.

1. Introduction

The aim of this paper is to develop and utilise a probabilistic model to aid with pedigree reconstruction, using (dominant-marker) phenotype data in the form of band absence/presence (from PCR amplification). There is a rich literature concerning genealogy estimation using genotype data, but considerably less attention has so far been paid to the (actual) case where only phenotype data is available. Estimating genealogy using phenotypes is considerably more involved (and computationally taxing) than its genotype counterpart for numerous reasons. Suppose we consider only the case (as we do in this paper) of estimation of parent–child relationships, and that birth data is available. Then a possible method of reconstruction using genotype data is the following: first split the individuals into founders and non-founders (if possible), and then sequentially for each non-founder and each pair of individuals with earlier birth dates, compute a likelihood that this pair is the parentage of the non-founder. A pedigree can be constructed using the maximum likelihoods of each non-founder. On the other hand, if only phenotype data is available, modifications have to be made to this algorithm. One possible approach is to first condition on the possible genotypes given the phenotypes and then compute the likelihood as a mixture of these conditional likelihoods (weighted by the distributions of the genotypes given the phenotypes). A problem with this approach is that one must constantly update the genotype distributions based on previous inferences. Another is that one mistake (early on in the inference) could lead to a pedigree differing greatly from the true pedigree (that one is trying to estimate). A possible workaround is to keep multiple options open throughout the inference, but this approach could result in an exponential growth (as the sequential reconstruction proceeds) of the number of possible pedigrees. Further discussion of these ideas can be found in Thompson (1976).

This is in fact not the approach we take in this paper (due to its high level of complexity, and the various problems that have to be resolved). Instead we assume that we have Hardy–Weinberg equilibrium (and we give a rigorous justification of this, under certain assumptions, in the appendix). Using this together with the assumption that gene-frequencies in the population are at equilibrium (i.e. do not vary over time), we are able to estimate various probabilities of interest that will enable the pedigree to be reconstructed. In particular, our approach avoids the need to continually update genotype distributions, but at the sacrifice of some assumptions on the population.

We work in this paper with a parent likelihood function. The use of such likelihoods can result in incorrect parent–child relationships to be inferred, particularly in the case in which siblings are present in the population. This somewhat paradoxical statement has been studied in Thompson and Meagher (1987) in which it is observed that sibling

☆ This work was supported by MODEMAVE research project from the Région Pays de la Loire. Access to molecular data was supported by both EUROGENI project, funded by Région Pays de la Loire (dynamiques de filière) and by BRIO project funded by same Région Pays de la Loire and the Fonds Unique Interministériel.

* Corresponding author.

http://dx.doi.org/10.1016/j.jtbi.2016.09.014
Received 22 February 2016; Received in revised form 29 August 2016; Accepted 6 September 2016
Available online 24 September 2016

relationships can give higher values than the true parents to the likelihood. They suggest that a further analysis could be carried out on individuals found to give a high value to the parent likelihood function, via a full sib likelihood function, which may enable a distinction to be made between parents and siblings of an individual. On the other hand, from a computational viewpoint, this problem may be too complex for large datasets with many individuals. Moreover, a suitable model which describes full sib phenotype relationships must be developed. Such a further analysis has been omitted from this work.

A similar model has been developed and used to construct maximum likelihoods in Jones (2003). Here, it is assumed that location-based data is available, so information about the locations of individuals can be incorporated into the model. One way we could include such data into our model would be via an adjustment of the prior distribution, i.e. by weighting the prior by a function of the distance between an individual and the possible parents.

Analysis of absence/presence of bands from DNA fingerprinting has been carried out in Jeffreys et al. (1991) and Geyer et al. (1993) and used to study relationships in a human and condor population, respectively. Deterministic methods based on the maximum parsimony principle and using purely combinatorial arguments allow a reconstruction of the minimal pedigree relating individuals in accordance with their types, see Chapter 4 in Semple and Steel (2003), Steel and Hein (2006) and Blouin (2003). There are also numerous different stochastic methods of reconstruction of pedigrees, see for instance Thatte (2013), Kirkpatrick et al. (2011), Thompson (2000), Thatte and Steel (2008), and Blouin (2003). Some of these models focus on the reconstruction of the lineages by estimating transition probabilities between nodes. Reconstructing the pedigree then comes down to the construction of a Markov chain. This method is quite popular when making use of identity by descent (IBD) data (Kirkpatrick et al., 2011). In this case, a statistical inference based on Monte Carlo Markov chains and Bayesian statistics are used to infer transition probabilities between nodes of the graph (Steel et al., 1998; Thompson, 2000). Coalescence theory may also prove to be a powerful tool in reconstruction of pedigrees, as observed in Wakeley et al. (2012).

1.1. Preliminaries

Mathematically, a pedigree is defined to be a directed acyclic graph with the property that each vertex has indegree equal to 0 or 2 and any outdegree (with further conditions in the case where individuals cannot be either male or female: see Thatte and Steel, 2008, Lemma 1). When it represents family relationships between individuals, the directions of edges indicate parent–child relationships, with the direction from parent to child. A pedigree is 'reconstructed' by determining, using a statistical analysis based on certain data, the most likely pedigree under a given model. If the model is correct then the pedigree of maximum likelihood should be a good candidate for the true pedigree. Possible data available could include phenotype, genotype, date of birth, data obtained from professional breeders, etc., and it may be that such data is only available for a subset of the individuals in the population.

In this paper, we work specifically with data provided through ISSR amplification for diploid plant cultivars, which are vegetatively propagated. ISSR amplification was popularised by Wolfe et al. (1998) and largely used in genetic diversity assessment (Pradeep Reddy et al., 2002). As a result of being vegetatively propagated, the data contains both descendants and ancestors in the pedigree, i.e. both terminal and internal nodes of the graph, while most above listed methods use information from last generation descendants (terminals). Our data provides the same information for each individual and one of the assumptions we make is that there are no missing individuals in the set. ISSR data only allows us to know, for each gene, if a specific allele is present or not. In the case of presence, we do not know if this specific allele is present in both chromosomes (i.e. at homozygotic state, and transmitted to all the descendants) or if it is present only in one of them (i.e. at heterozygotic state and thus transmitted to only half of the descendants).

Since much of the discussion is for a general dataset, we introduce some notation. We shall denote by n the number of individuals for which data is available, and write $g_1, \ldots, g_n$ for these individuals ranked in their birth order (from oldest to youngest). This set of individuals is partitioned into founders, denoted F, and non-founders, denoted F^c. We shall write m for the number of (binary) data points available for each individual (noting that, on the assumption of no missing data, this value is the same for each individual). Note also that this does not include date of birth data. Then the aim is to find, for each $i = 1, \ldots, n$ corresponding to a non-founder g_i, a pair of individuals (possibly non-distinct) from the set $\{g_1, \ldots, g_{i-1}\}$ which (under a given model) is the most likely parent pair of g_i. The construction of this model hinges on the fact that the probability to observe the specific alleles in the child, given the status of the parents, does not depend on the generation. It only depends on the gene frequencies which we assume are constant in time. In order to justify this assumption, we shall prove in an appendix that gene frequencies converge almost surely, as the number of crossbreeding increases, toward an equilibrium which satisfies the Hardy–Weinberg condition.

Our reconstruction method is applied to a family of 85 living accessions representing the common broom *Cytisus scoparius* and related cultivated hybrids (*Cytisus* x *dallimorei* and *Cytisus* x *boskoopi*). The latter are diploid sexed plants whose crossbreedings have occurred in the past 200 years from a set of founders. For each individual, 6 markers are used to highlight presence or absence of a particular allele in a high number of distinct regions of the genome. These 6 markers provide a total of more than 420 distinct bands for these 85 accessions, and each band has been treated as present or absent for each individual. The results of our model applied to these particular data are described in Section 3. Section 2 is devoted to the presentation of the model. We give some conclusions in Section 4, comparing our results to the existing literature and highlighting some other frameworks where our method can be used.

2. Materials and methods

2.1. Model overview

We begin with a discussion about the various assumptions on the individuals and data necessary to construct our probabilistic model. The first aspect of this is an understanding of how reproduction occurs within the population. The m binary data points for us will refer to the presence (1) or absence (0) of an allele. Indeed, with ISSR amplification using a particular marker, a binary response of 1 indicates that the allele is present in at least one of the two chromosomes and a response of 0 indicates it is absent in both. In particular, when the allele is present, we do not know if it is present on the two chromosomes.

For each individual g_i and each gene $\ell \in \{1, \ldots, m\}$, we let $x_\ell(g_i) \in \{0, 1\}$ be the indicator of absence and presences of individual g_i obtained during the ISSR amplification process. Hence the *apparent genotype* of each individual g will be identified with the vector $x(g) := (x_1(g), x_2(g), \ldots, x_m(g)) \in \{0, 1\}^m$. Note that the event $\{x_\ell(g) = 1\}$ means 'one observes the presence of the allele specific to gene ℓ in individual g' or equivalently 'the allelic combination of gene ℓ in individual g is 01 or 11'.

As mentioned, in addition to the m binary data points, each individual g has an associated date of birth, which we denote $t(g)$. Recall that the individuals are ordered, and that this ordering satisfies if $i < j$ then $t(g_i) \leq t(g_j)$. In determining possible parent pairs, it is in fact only the orders of the various dates that matter and so the values of $t(g)$ will be taken to be non-negative, with founders having value 0. In our particular dataset, several of the individual are obtained from the wild and precisely these individuals will be considered founders. Our model assumes that little is known a priori about the relationships

between the individuals other than their relative birth dates, and we formalise this with a uniform prior on the probability that (g_j, g_k) are the parents of individual g_i over all pairs (g_j, g_k) with $\max(t(g_j), t(g_k)) < t(g_i)$. We also that there are no missing individuals, so that the parents of each non-founder individual g_i belong to the set $\{g_1, \ldots, g_n\} \setminus \{g_i\}$. To compensate for this fairly strong assumption, we will impose a *threshold probability* and reject the most likely parent pair of an individual if its corresponding likelihood falls below this threshold level. Since this level is imposed artificially, we shall present several pedigree reconstructions for differing values of the threshold. Note that there is a monotonicity property: decreasing the threshold cannot remove any edges from the reconstructed pedigree.

A further assumption we make is that for each individual g, the coordinates of the apparent genotype are pairwise independent, that is, for every $\ell \neq \ell' \in \{1, \ldots, m\}$, events $\{x_\ell(g) = 1\}$ and $\{x_{\ell'}(g) = 1\}$ are independent. In other words, we assume that alleles are 'independently inherited'. As we shall see, this assumption allows the likelihood function to exhibit a product form, which greatly reduces the algorithmic complexity.

We introduce some more notation to describe the distribution of the apparent genotype of offspring, given the apparent genotype of the parents. Let $\delta \in (-1/2, 1/2)$ and $\varepsilon \in (0, 1/2)$ be constants which we will refer to as *errors*. These compensate for any experimental errors (i.e. in a laboratory). For each ℓ, we let $p_\ell \in (3/4, 1)$ and $q_\ell \in (1/2, 1)$ be constants. These constants satisfy the following properties: for each individual g with parents $\hat{g}$ and $\overline{g}$, we have

- $\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 1\}, \{x_\ell(\overline{g}) = 1\}) = \min(p_\ell - \delta, 1)$,
- $\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 0\}, \{x_\ell(\overline{g}) = 1\}) = \min(q_\ell - \delta, 1)$,
- $\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 0\}, \{x_\ell(\overline{g}) = 0\}) = \varepsilon$.

The values of these constants must be estimated using the data available. We detail how this can be done in the next section.

Finally we remark that, despite the reproduction being sexual, since we are concerned in this work with plant populations, each individual can either be male or female. Thus when referring to the parents g_j and g_k of the individual g_i the mother and the father are not distinguished. In particular we have

$$\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 0\}, \{x_\ell(\overline{g}) = 1\}) = \mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 1\}, \{x_\ell(\overline{g}) = 0\}).$$

2.2. Estimating model parameters

We focus in this section on the estimation of the values p_ℓ and q_ℓ. To begin with we assume that there is no experimental error, i.e. $\delta = \varepsilon = 0$, so that the expressions in the previous section become

$$\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 1\}, \{x_\ell(\overline{g}) = 1\}) = p_\ell,$$

and

$$\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 0\}, \{x_\ell(\overline{g}) = 1\}) = q_\ell.$$

These constants can be estimated by observing the various frequencies of the three possible genotypes (00, 01 and 11) appearing in the dataset. However, since we cannot distinguish between 01 and 11, we must estimate p_ℓ and q_ℓ using only the frequency of 00. We detail how this is possible once we assume that these frequencies are at an equilibrium. We denote by $\pi_{00}(\ell)$, $\pi_{01}(\ell)$ and $\pi_{11}(\ell)$ these frequencies and assume that equilibrium is attained, so that $\pi_{00}(\ell)$, $\pi_{01}(\ell)$ and $\pi_{11}(\ell)$ do not change over time (successive crossbreedings). We justify this assumption rigourously in the appendix.

Lemma 1. *For each ℓ assume we have Hardy–Weinberg equilibrium*: $\pi_{01}(\ell) = 2\sqrt{\pi_{00}(\ell)\pi_{11}(\ell)}$. *Then*

$$q_\ell = \frac{1}{1 + \sqrt{\pi_{00}(\ell)}}, \quad p_\ell = \frac{1 + 2\sqrt{\pi_{00}(\ell)}}{(1 + \sqrt{\pi_{00}(\ell)})^2}. \tag{2.1}$$

Proof. In the following, when no confusion is possible, we will (for sake of simplicity of presentation) drop the index ℓ in $\pi_{00}(\ell)$, $\pi_{01}(\ell)$ and $\pi_{11}(\ell)$. We first compute p_ℓ and q_ℓ in terms of π_{00}, π_{11} and π_{01}. For a fixed individual g, and a pair of individuals $(\hat{g}, \overline{g})$ each chosen uniformly at random from the sub-population $\{g' : t(g') < t(g)\}$, the probability to observe $x_\ell(\hat{g}) = 1$ and $x_\ell(\overline{g}) = 1$ is

$$\mathbb{P}(\{x_\ell(\hat{g}) = 1\}, \{x_\ell(\overline{g}) = 1\}) = \pi_{11}^2 + 2\pi_{01}\pi_{11} + \pi_{01}^2.$$

When they breed and give a child g, we have

$$\mathbb{P}(\{x_\ell(g) = 1\}, \{x_\ell(\hat{g}) = 1\}, \{x_\ell(\overline{g}) = 1\}) = \pi_{11}^2 + 2\pi_{01}\pi_{11} + 3\pi_{01}^2/4.$$

Therefore p_ℓ satisfies

$$p_\ell = \frac{\pi_{11}^2 + 2\pi_{01}\pi_{11} + 3\pi_{01}^2/4}{\pi_{11}^2 + 2\pi_{01}\pi_{11} + \pi_{01}^2} = 1 - \frac{\pi_{01}^2}{4(\pi_{01} + \pi_{11})^2},$$

and q_ℓ can be obtained in the same way:

$$q_\ell = \frac{\pi_{01} + 2\pi_{11}}{2\pi_{01} + 2\pi_{11}}.$$

Since the frequencies π_{00}, π_{01} and π_{11} belong to $(0, 1)$, it is easy to check from the above expressions that $p_\ell \in (3/4, 1)$ and $q_\ell \in (1/2, 1)$. Furthermore, we have the relationship $p_\ell = q_\ell(2 - q_\ell)$.

Now using the relation $\pi_{01} = 2\sqrt{\pi_{00}\pi_{11}}$, we deduce that

$$q_\ell = \frac{1}{1 + \sqrt{\pi_{00}}}, \qquad p_\ell = \frac{1 + 2\sqrt{\pi_{00}}}{(1 + \sqrt{\pi_{00}})^2}, \tag{2.2}$$

as claimed.□

2.3. Model construction

We shall now define the set of probability measures from which the most likely pedigree will be derived. This definition is based on the following product form of the conditional probabilities:

$$\mathbb{P}(x(g) = a | x(\hat{g}) = \hat{a}, x(\overline{g}) = \overline{a}) = \prod_{\ell=1}^{m} \mathbb{P}(x_\ell(g) = a_\ell | x_\ell(\hat{g}) = \hat{a}_\ell, x_\ell(\overline{g}) = \overline{a}_\ell),$$

which are obtained from all possible triplets of individuals $(g, \hat{g}, \overline{g})$ and their apparent genotypes $a = (a_1, \ldots, a_m)$, $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_m)$ and $\overline{a} = (\overline{a}_1, \ldots, \overline{a}_m)$ in $\{0, 1\}^m$. More specifically, the set of individuals $\{g_1, \ldots, g_n\}$ and their apparent genotype being given, for all triples $(i, j, k) \in \{1, \ldots, n\}^3$ and for each $\ell \in \{1, \ldots, m\}$, we define the agreements/disagreements indicators between the genotype of an individual g_i and that of the possible pair of parents (g_j, g_k):

$$p_{ijk}^{(\ell)} = \mathbf{1}_{\{x_\ell(g_j) = x_\ell(g_k) = x_\ell(g_i) = 1\}}, \quad \overline{p}_{ijk}^{(\ell)} = \mathbf{1}_{\{x_\ell(g_j) = x_\ell(g_k) = 1, x_\ell(g_i) = 0\}},$$

$$q_{ijk}^{(\ell)} = \mathbf{1}_{\{x_\ell(g_j) \neq x_\ell(g_k), x_\ell(g_i) = 1\}},$$

$$\overline{q}_{ijk}^{(\ell)} = \mathbf{1}_{\{x_\ell(g_j) \neq x_\ell(g_k), x_\ell(g_i) = 0\}} \varepsilon_{ijk} = \sum_{\ell=1}^{m} \mathbf{1}_{\{x_\ell(g_j) = x_\ell(g_k) = 0, x_\ell(g_i) = 1\}},$$

$$\overline{\varepsilon}_{ijk} = \sum_{\ell=1}^{m} \mathbf{1}_{\{x_\ell(g_j) = x_\ell(g_k) = x_\ell(g_i) = 0\}}.$$

Now set $p_{\delta,\ell} = \min(p_\ell - \delta, 1)$, $q_{\delta,\ell} = \min(q_\ell - \delta, 1)$, $\overline{p}_{\delta,\ell} = 1 - p_{\delta,\ell}$, $\overline{q}_{\delta,\ell} = 1 - q_{\delta,\ell}$, $\overline{\varepsilon} = 1 - \varepsilon$ and for $j \leq k < i$, define

$$\log \nu_i(j, k) := \varepsilon_{ijk} \log \varepsilon + \overline{\varepsilon}_{ijk} \log \overline{\varepsilon} + \sum_{\ell=1}^{m} \{p_{ijk}^{(\ell)} \log p_{\delta,\ell} + \overline{p}_{ijk}^{(\ell)} \log \overline{p}_{\delta,\ell} + q_{ijk}^{(\ell)} \log q_{\delta,\ell} + \overline{q}_{ijk}^{(\ell)} \log \overline{q}_{\delta,\ell}\},$$

and otherwise set $\nu_i(j, k) = 0$. Then for each non-founder g_i, the probability measure μ_i on $\{1,\ldots,n\}^2$ is explicitly defined in terms of $x(g_i)$ as

$$\mu_i(j, k) = \frac{\nu_i(j, k)}{z_i}, \quad j, k \in \{1,\ldots,n\},$$

where $z_i := \sum_{j,k} \nu_i(j, k)$ is a normalising constant. We readily check that $z_i > 0$ for all i such that $t(g_i) > 0$.

Writing $p \in (0, 1)$ for the threshold probability, we reconstruct the pedigree by determining, for each $g_i \in F^{\mathbb{C}}$, the individuals g_j and g_k (possibly equal), satisfying:

1. $j \le k < i$;
2. $\mu_i(j, k) = \max_{j',k'}\{\mu_i(j', k')\colon j' \le k' < i\}$ (g_j and g_k maximise the likelihood);
3. $\mu_i(j, k) \ge p$.

We remark that the normalisation of the probability measure μ is relevant only for the comparison with the threshold probability. The pedigree reconstructions displayed in Section 3 have been obtained by implementing the above steps with an R program for our specific data. Note that the implementation requires an estimation of the values of ε and δ (or else, they can be set to 0).

3. Application of the model

Our particular dataset concerns a population of 85 living accessions representing the common broom *C. scoparius* and three related interspecific hybrids. This dataset consists of 62 vegetatively propagated cultivars obtained from various nurseries. These cultivars belong to either *C. scoparius*, *Cytisus* x *dallimorei* (hybrid between *C. scoparius* and *C. multiflorus*), *C.* x *praecox* (hybrid between *C. multiflorus* and *C. oromediterraneus*), or *C.* x *booskopii* (hybrid between *C.* x *dallimorei* and *C.* x *praecox*). In addition three to nine individuals obtained from each of five wild populations have been included (3 individuals of *Cytisus oromediterraneus* from France, 3 individuals of *C. scoparius* from Italia, 3 from Poland, 4 from Angers, France and 9 from Ernée, France). For all these samples, DNA extraction used the Nucleospin® Plant II kit from macherey-Nagel. IISR data was obtained using six set of primers, namely ISSR5 (sequence: 5-CACAC ACACACACACARC-3), ISSR7 (sequence: 5-CACACACACACACAC ART-3), ISSR13 (sequence: 5-GTGTGTGTGTGTGTGTYA-3), ISSR890 (sequence: 5-VHVGTGTGTGTGTGTGT-3), ISSR891 (sequence: 5-HVHTGTGTGTGTGTGTG-3) and ISSRa (sequence: 5-GCTCTCTC TCTCTCTC-3). Polymerase chain reaction (PCR) was done using the following parameters : 95 °C for 2 min, then 39 cycles of 95°C for 30 s, 50 °C for 30 s, 72 °C for 120 s, followed by 10 min of extension at 72 °C. Electrophoresis was done on 5% acrylamide–bisacrylamide gel (mixing ratio: 29:1), with 7 M urea, with a pre-run of 30 min at 80 W, then 2 h30 at 60 W. Staining used silver nitrate. Gels were scanned and band read manually.

Using the data obtained in this manner, our aim in this section is to determine the most likely pedigree relating these individuals. A code in language R has been written according to the model described in the previous sections. This code applied to our data provided the pedigrees presented in Figs. 1–3. The parameters ε, δ, p_ℓ and q_ℓ must be inferred from our data in order to use the model outlined in the previous sections.

Breedings have occurred over time under the action of professional breeders or according to natural phenomena and with no more information available, our uniform prior assumption is reasonable. On the other hand, some relationships between certain individuals are believed to hold, and we will discuss shortly how our model (taking a uniform prior) is able to recover many of these relationships.

We believe that for our particular dataset few or no individuals are missing, and so if no parents can be found for an individual then we will assume that individual is a founder. We next need to justify the independence between the bands $\{x_\ell(g) = 1\}$, $\ell \in \{1,\ldots,m\}$. However, dependence may occur due to the selective sweep phenomenon which can associate together several genes whose loci are close to each other along the chromosome. For such sets of genes, recombination is not strong enough for them to be considered as independent in the reproduction process. As an attempt to rectify this we have manually selected 168 of the 424 bands which we will assume satisfy the independence assumption. These 168 bands are split across the four markers as 34 bands for ISSR890, 22 for ISSR891, 31 for ISSRa, 32 for ISSR5, 27 for ISSR7 and 22 for ISSR13. The choice of selection is based purely on an observation of joint presence and absence of bands across

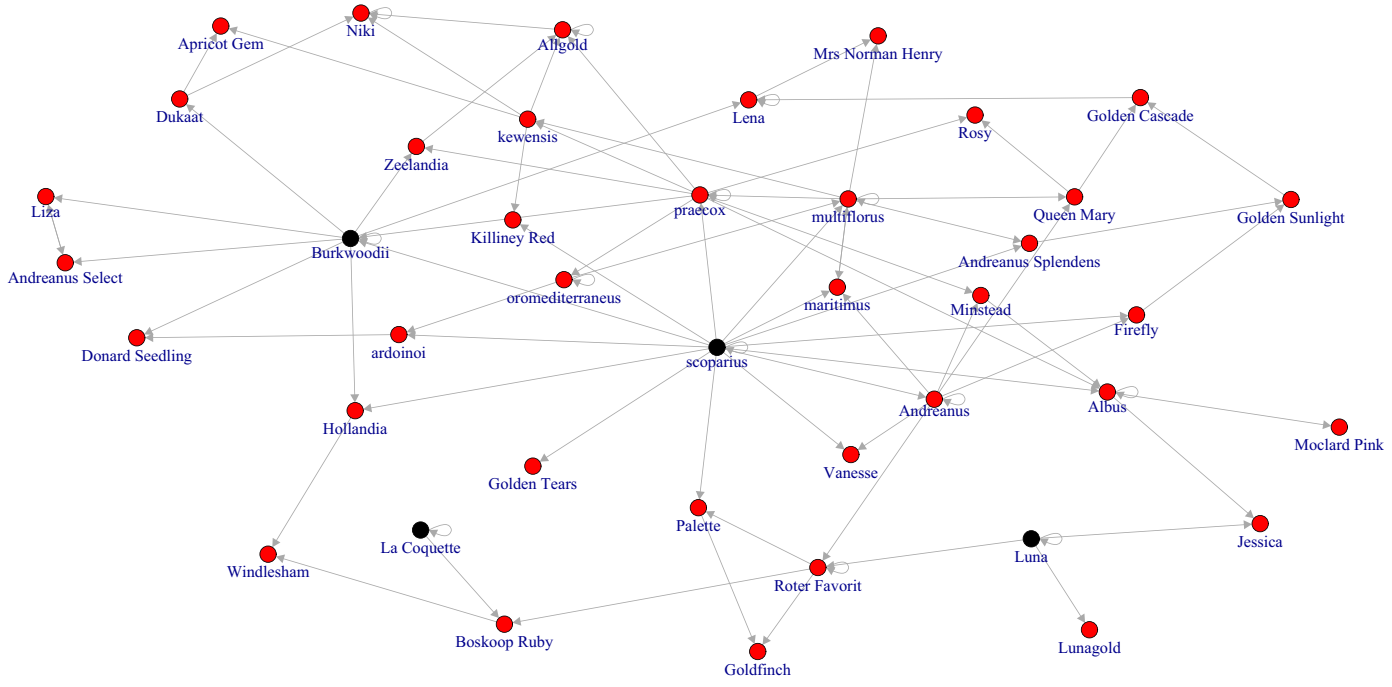

Fig. 1. Threshold probability p=0.1.

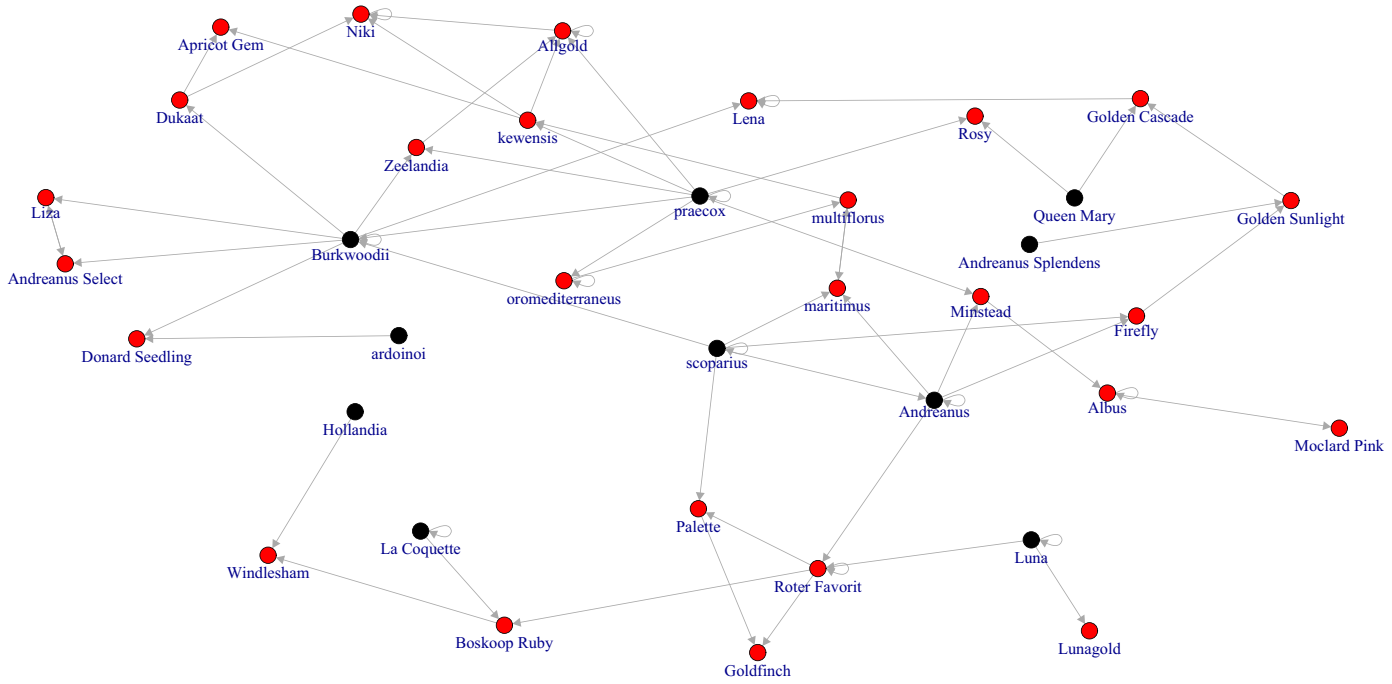

Fig. 2. Threshold probability p=0.2.

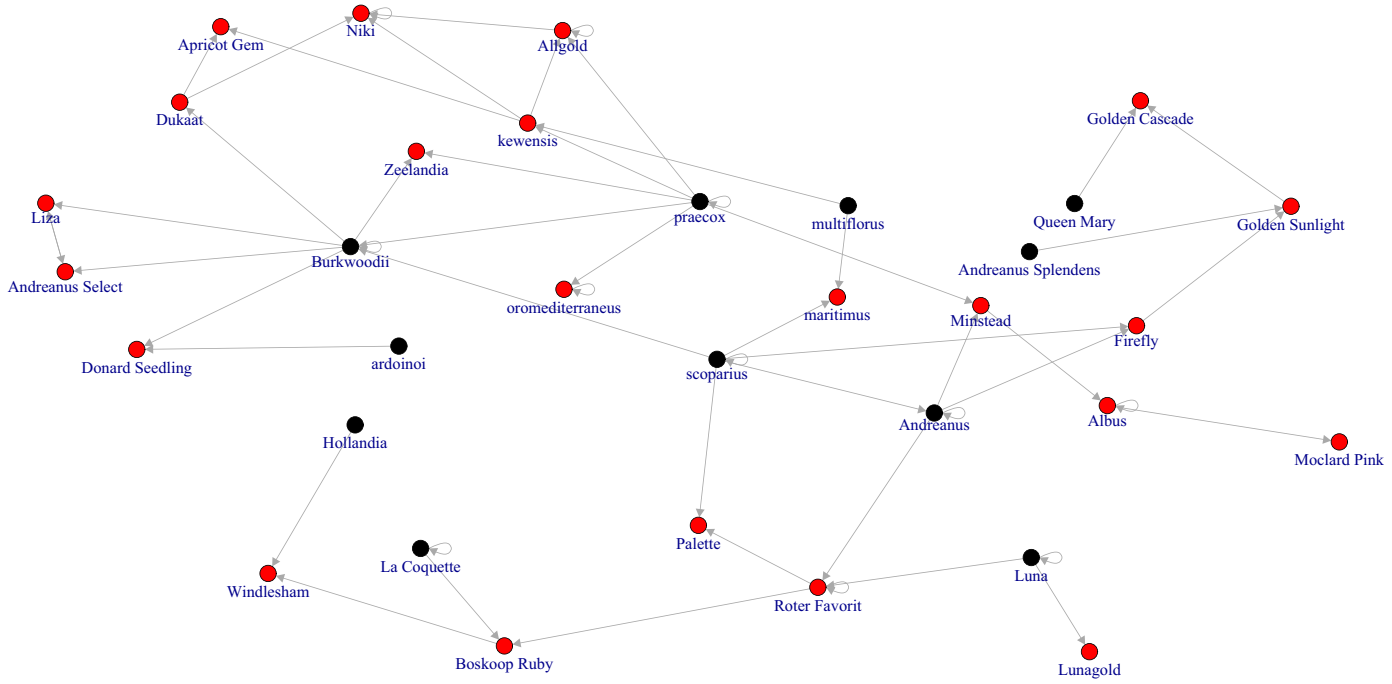

Fig. 3. Threshold probability p=0.3.

the population. More specifically, we place bands into equivalence classes in the following manner: for each pair of bands within a marker, we observe the number of present–present, absent–present and absent–absent occurrences across all individuals. If, based on these numbers, we consider the two bands to be strongly correlated, we say they are in the same equivalence class. To obtain our 'independent' bands, we will simply choose a single member from each equivalence class (i.e. a class representative).

We also need to determine the values of ε, δ, p_ℓ and q_ℓ related to the present data, in order to construct the probability measure defined in the model. We achieve this by repeatedly crossing two individuals (G017 *C. scoparius* '*Lunagold*' and G010 *Cytisus* x *dallimorei* '*Burkwoodii*') and performing marker analysis (the data provided to us for these crossbreedings used only 5 of the 6 ISSR markers used for the full dataset) on the resulting offspring (n=33 plants). No knowledge of p_ℓ or q_ℓ (for the crossbreeding dataset) is required to estimate the value of ε. On the other hand, to estimate δ, we must first estimate the values of p_ℓ and q_ℓ (found via π_{00} and Hardy–Weinberg principle). We obtain the following estimates (by taking the mean values obtained over all markers): $\delta = 0.15$ and $\varepsilon = 0.05$.

In the appendix we prove convergence of gene frequencies and we will assume that the population which is considered here has attained some equilibrium. As can be seen from Eq. (2.2), thanks to Hardy–Weinberg principle, the probabilities p_ℓ and q_ℓ only depend on the

probability π_{00}. We emphasise that the latter probability is actually the only one whose empirical value can be determined from the data. Indeed it is not possible to distinguish the genotype 01 from the genotype 11 in ISSR data. In the present case, we obtain the values of π_{00} and hence p_ℓ and q_ℓ for each band.

The probabilities $\mu_i(j, k)$ defined in the end of Section 2.1 may appear quite low once computed from our dataset. However knowing that all individuals belong to the same family, we are only concerned with their relative values. The pedigrees appearing in Figs. 1, 2 and 3 were obtained with the threshold probabilities 0.1 and 0.2 and 0.3 respectively. Individuals whose parentage cannot be determined and founders have been represented in black and individuals have been omitted if no parents or children are found by the model. As expected, when the threshold probability p increases, the number of relations between individuals decreases and more individuals are considered as founders. Compared to the existing knowledge we have on the group (see Auvray, 2011), several relationships are congruent with historical information. For example, 'Zeelandia' is reported as a descendant of 'Burkwoodii' and a *C.* x *praecox*. This relationship appears with all threshold probabilities. 'Liza', 'Andreanus Select', and 'Donard Seedling' are all historically reported as sport (bud mutations) of 'Burkwoodii', while 'Lena' is supposed to be a seedling of it. They are all linked under p=0.1 and p=0.2, while under higher threshold probability 'Burkwoodii', 'Liza' and 'Andreanus Select' are still linked, however, Donard Seedling is treated as a seedling of 'Burkwoodii' and Cytisus ardoinoi which may be impossible (the sample used for representing this last species being wild collected). 'Firefly' is reported as a seedling of 'Andreanus', which appears under all threshold probabilities. Comparing to historical information, 'La Coquette' appears here as founder, and as parent of 'Roter Favorit' while it was reported as a self-fecondation of 'Hollandia', and half-brother of 'Boskoop Ruby'. 'Hollandia' is known to be a seedling from 'Burkwoodii' and *C.* x *praecox*, here, under p=0.1, it is a seedling between the same 'Burkwoodii' but with *C. scoparius*.

Using the same ISSR data, Auvray in Auvray (2011) points out the putative link between 'Apricot Gem' and 'Dukaat', as well as between 'Boskoop Ruby' and 'Windlesham'. These links are re-inforced here and second putative parents are provided (kewensis for 'Apricot Gem' and 'Hollandia' for 'Windlesham'). Auvray (2011) also point out a parentage between 'Moclard Pink' and 'Minstead' (the former being a putative seedling of the later), here 'Moclard Pink' is always linked with 'Albus', a point which needs consideration. Under the various threshold probabilities, 'Luna', 'Palette' and 'Roter Favorit' are linked, this seems reasonably consistent with the fact that they all have been obtained from the same nursery (Arnold, at Alreslohe near Holstein in Germany) around 1960. 'Jessica', linked to the same group under p=0.1 is of unknown parentage, while 'Goldfinch', also linked under p=0.1 is reported to be a seedling between 'Donard Seedling' and 'Dorothy Walpole' (lacking from the sampling). The links between 'Andreanus', 'Firefly', 'Golden Sunlight', 'Andreanus Splendens', 'Golden Cascade', 'Roter Favorit' and 'Queen Mary', appearing under all threshold probabilities, is a reminder that all these cultivars are selection of *C. scoparius* and not of any of the interspecific hybrids.

4. Discussion

We have set up a mathematical model of pedigree reconstruction whose basic principle is to determine, for each individual, what is the most likely parent pair in the population, according to a certain probabilistic model. The robustness of this model mainly relies on the fact that gene frequencies have attained some equilibrium. We have shown (see the appendix) that indeed, in the absence of any evolutive forces, gene frequencies converge toward a limit random vector which satisfies Hardy–Weinberg equilibrium. From this model we derived an algorithm which is written in language R and then we applied this model to ISSR data from a population of diploid plants. The results reveal that the pedigrees obtained from this method fit to the partial reconstructions based on botanical data or other methods using dendograms obtained from matrix distances. This additional source of information could also be used in order to improve the model by constructing a new probability distribution giving a relative weight to each kind of data.

We have assumed that recombination is uniform, but this could be made more realistic by determining how different sets of loci actually recombine from a preliminary statistical inference. The model could easily be adapted to this setting.

Finally we emphasise that our model has essentially been applied to phenotype data. Indeed, as already observed in Section 2, the knowledge of ISSR is equivalent to the knowledge of the expression of a dominant gene. Hence our model can easily be tested from a population about which we observe a specific set of phenotypical criteria and whose family relationship are a priori known.

Acknowledgements

Projects EUROGENI and BRIO have been managed by Véronique Kapusta, while molecular and bibliographic information concerning Cytisus material had been acquired by Gaëlle Auvray, Agathe Le Gloanic and Nadège Le Pocreau. We warmly thank all of them for their help.

Appendix A

We show that the assumption of Lemma 1 is satisfied, that is, the frequencies π_{00}, π_{01} and π_{11} of the types 00, 01 and 11 can be taken to be fixed in time, and satisfy the Hardy–Weinberg condition.

From time n=0, we rank the crossbreedings in increasing order as they occur. Since we assume the evolution of genes are independent of each other, we only need to consider the dynamics of the frequencies of genotypes 00, 01, 11 for one gene. Suppose π^n_{00}, π^n_{01} and π^n_{11} denote the proportions of individuals g with genotype 00, 01 or 11 respectively, just after the n-th crossbreeding. We assume that we start at time n=0 with two founders, so that after the n-th crossbreeding, $n + 2$ individuals are present in the population. In particular, there is no death which is consistent with the fact that we consider plant cultivars in this work. Moreover we assume that both alleles exist in the two founders. Then our reproduction law described in the previous sections may actually be represented as a generalised urn model in which the probability of replacement depends on the proportion of individuals in the population, see Pemantle (2007) and the references therein. More specifically, at each step n, we choose two individuals uniformly at random in the population.

Consider the polynomial function F: $\{(x, y, z) \in [0, 1]^3 : x + y + z = 1\} \to \mathbb{R}^3$ given by

$$F(x, y, z) + (x, y, z) = (xy + x^2 + y^2/4,\ xy + yz + 2xz + y^2/2,\ yz + z^2 + y^2/4),$$

and denote by $\mathcal{S} = \{(x, y, z) \in [0, 1]^3 : F(x, y, z) = 0\}$ the zero set of F.

We construct $\pi^n = (\pi^n_{00}, \pi^n_{01}, \pi^n_{11})$ recursively. Write $F = (F_1, F_2, F_3)$. At each step n, two uniformly chosen individuals from the population breed

and the new frequencies of individuals with types 00, 01 and 11 become:

$$\begin{cases} \pi_{00}^{n+1} = \dfrac{(n+2)\pi_{00}^n + 1}{n+3} \\ \pi_{01}^{n+1} = \dfrac{(n+2)\pi_{01}^n}{n+3}, \\ \pi_{11}^{n+1} = \dfrac{(n+2)\pi_{11}^n}{n+3} \end{cases} \text{ with probability } \pi_{00}^n\pi_{01}^n + (\pi_{00}^n)^2 + (\pi_{01}^n)^2/4 = F_1(\pi^n),$$

$$\begin{cases} \pi_{00}^{n+1} = \dfrac{(n+2)\pi_{00}^n}{n+3} \\ \pi_{01}^{n+1} = \dfrac{(n+2)\pi_{01}^n + 1}{n+3}, \\ \pi_{11}^{n+1} = \dfrac{(n+2)\pi_{11}^n}{n+3} \end{cases} \text{with probability } \pi_{00}^n\pi_{01}^n + \pi_{01}^n\pi_{11}^n + 2\pi_{00}\pi_{11} + (\pi_{01}^n)^2/2 = F_2(\pi^n),$$

$$\begin{cases} \pi_{00}^{n+1} = \dfrac{(n+2)\pi_{00}^n}{n+3} \\ \pi_{01}^{n+1} = \dfrac{(n+2)\pi_{01}^n}{n+3}, \\ \pi_{11}^{n+1} = \dfrac{(n+2)\pi_{11}^n + 1}{n+3} \end{cases} \text{ with probability } \pi_{01}^n\pi_{11}^n + (\pi_{11}^n)^2 + (\pi_{01}^n)^2/4 = F_3(\pi^n),$$

Let us make this construction more formal. First we define a stochastic process $(\delta_n)_n$ with values in $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ in such a way that the law of δ_{n+1} conditionally on $\pi^0 = i_0, \ldots, \pi^n = i_n$ is $F(i_n)$. Recall that the quantity $(n+2)\pi^n$ is the vector of the numbers of individuals of type 00,01,11 at time n. Then π^{n+1} is defined by

$$(n+3)\pi^{n+1} = (n+2)\pi^n + \delta_{n+1}, \quad n \geq 0.$$

Let us set

$$\eta_n = \delta_{n+1} - F(\pi^n),$$

then we readily obtain the following equality:

$$\pi^{n+1} = \pi^n + \frac{1}{n+3}(F(\pi^n) - \pi^n + \eta_n). \tag{1.3}$$

For $u \in [0, 1]^3$, let $f_u : \mathbb{R}^+ \cup \{0\} \to [0, 1]^3$ be the solution to the ODE

$$\begin{cases} \dfrac{d}{dt} f_u(t) &= F(f_u(t)), \quad t \geq 0, \\ f_u(0) &= u. \end{cases} \tag{1.4}$$

The solution can be calculated explicitly and we easily check that with $f_u(t) = (x_u(t), y_u(t), z_u(t))$ and $u = (x_0, y_0, z_0)$, then

$$\begin{cases} x_u(t) = \left(x_0 - \dfrac{(2x_0 + y_0)^2}{4}\right)e^{-t} + \dfrac{(2x_0 + y_0)^2}{4} \\ y_u(t) = -2\left(x_0 - \dfrac{(2x_0 + y_0)^2}{4}\right)e^{-t} - \dfrac{(2x_0 + y_0)^2}{2} + 2x_0 + y_0 \\ z_u(t) = 1 + \left(x_0 - \dfrac{(2x_0 + y_0)^2}{4}\right)e^{-t} + \dfrac{(2x_0 + y_0)^2}{4} - 2x_0 - y_0. \end{cases}$$

We aim to show almost-sure convergence of π^n as $n \to \infty$. The first step in achieving this is to show almost-sure convergence of $v(\pi^n)$ as $n \to \infty$, where $v(u) := \lim_{t\to\infty} f_u(t)$. This is achieved in the following lemma.

Lemma 2. *As $n \to \infty$, $v(\pi^n)$ converges almost surely.*

Proof. We shall show that almost surely, $(v(\pi^n))_n$ is a Cauchy sequence. We have

$$|v(\pi^{n+1}) - v(\pi^n)| \leq \left| v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right) - v(\pi^n) \right| + \left| v(\pi^{n+1}) - v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right) \right|. \tag{1.5}$$

We provide upper bounds on each term appearing on the right-hand side. Firstly, using the fact that $v(x) = v(f_x(t))$ for any $t \geq 0$,

$$\left| v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right) - v(\pi^n) \right| = \left| v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right) - v\left(f_{\pi^n}\left(\frac{1}{n+3}\right)\right) \right|.$$

We have the explicit form of v as

$$v(u) = \left(\frac{(2x_0 + y_0)^2}{4}, -\frac{(2x_0 + y_0)^2}{2} + 2x_0 + y_0, 1 + \frac{(2x_0 + y_0)^2}{4} - 2x_0 - y_0\right),$$

for any $u = (x_0, y_0, z_0)$. The function v is clearly Lipschitz on $[0, 1]^3$ and so there exists a constant c such that

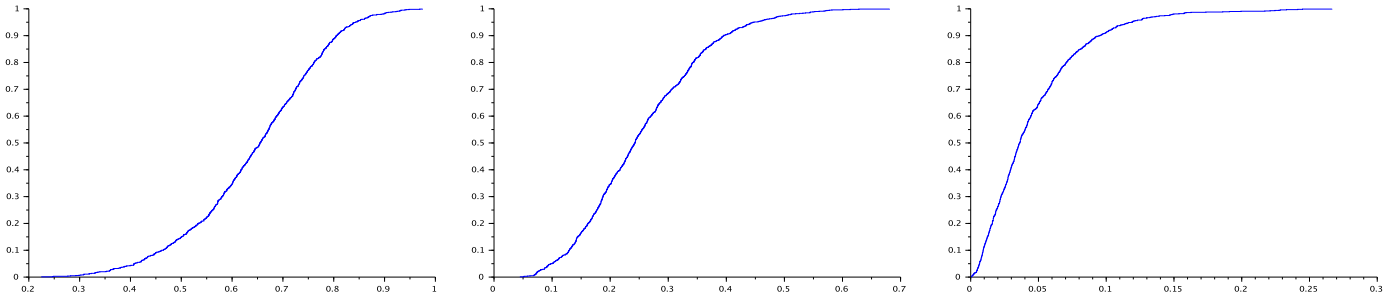

Fig. 4. Empirical distribution function of π_{00}. From left to right, figures are obtained respectively with initial values $(\pi_{00}^0, \pi_{01}^0, \pi_{11}^0) = (7/10, 1/5, 1/10)$; $(\pi_{00}^0, \pi_{01}^0, \pi_{11}^0) = (1/5, 3/5, 1/5)$ and $(\pi_{00}^0, \pi_{01}^0, \pi_{11}^0) = (1/10, 1/5, 7/10)$.

$$\left| v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right) - v\left(f_{\pi^n}\left(\frac{1}{n+3}\right)\right)\right| \leq c\left|\pi^n + \frac{1}{n+3}F(\pi^n) - f_{\pi^n}\left(\frac{1}{n+3}\right)\right| \leq O(1/n^2),$$

since $f_{\pi^n}(1/(n+3)) = f_{\pi^n}(0) + \frac{1}{n+3}f_{\pi^n}{}'(0) + O(1/n^2) = \pi^n + \frac{1}{n+3}F(\pi^n) + O(1/n^2)$. For the second term on the right-hand side of (1.5), we have

$$\left| v(\pi^{n+1}) - v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right)\right| \leq c\left|\pi^{n+1} - \pi^n - \frac{1}{n+3}F(\pi^n)\right| \leq \frac{c}{n+3}|\eta_n - \pi^n|,$$

by the definition of π^{n}, see (1.3). However since F is bounded we deduce that we can upper bound this term by $O(1/n)$. Plugging the two bounds into Eq. (1.5) shows that the sequence $(v(\pi^n))_n$ is indeed Cauchy (surely), and this completes the proof.□

We are now in a position to show almost-sure convergence of the stochastic process $\pi^n = (\pi_{00}^n, \pi_{01}^n, \pi_{11}^n)$, $n \geq 1$.

Theorem 1. *The random vector $\pi^n = (\pi_{00}^n, \pi_{01}^n, \pi_{11}^n)$, $n \geq 1$ has the following asymptotic behaviour*:

$$\pi^n \overset{a.s.}{\longrightarrow} (\pi_{00}, \pi_{01}, \pi_{11}), \quad \text{as } n \text{ tends to } +\infty,$$

where $(\pi_{00}, \pi_{01}, \pi_{11})$ is distributed on $\mathcal{S}$. In particular, it satisfies the Hardy–Weinberg equilibrium:

$$\pi_{01} = 2\sqrt{\pi_{00}\pi_{11}}.$$

Proof. We first claim that almost surely, the L^1 distance between π^{n} and $\mathcal{S}$ tends to 0 as $n \to \infty$. Recall that the L^1 distance $|\pi^n - \mathcal{S}|$ is defined as

$$|\pi^n - \mathcal{S}| := \min_{s\in\mathcal{S}}\{|\pi^n - s|\} := \min_{(x,y,z)\in\mathcal{S}}\{|\pi_{00}^n - x| + |\pi_{01}^n - y| + |\pi_{11}^n - z|\}.$$

In fact, this is a consequence of Theorem 2.2 in Schreiber (2001) which asserts that the limit set of (π^n) (i.e. the set of limits of subsequences of (π^n)) is almost surely a connected compact internally chain recurrent set for the flow associated to the ODE (1.4). In particular the limit set of (π^n) is included in $\mathcal{S}$, which implies that the distance between π^{n} and $\mathcal{S}$ tends almost surely to 0.

Suppose $x \in \mathcal{S}$ so that $F(x) = 0$ by definition. Then $\frac{d}{dt}f_x(t) = 0$ for all $t \geq 0$ and so $f_x(t) = x$ for all $t \geq 0$, and in particular $v(x) = x$. Since v is Lipschitz and $v(\mathcal{S}) = \mathcal{S}$ we have that, as $y \to \mathcal{S}$, $|v(y) - y| \to 0$. But since $v(\pi^n)$ converges almost surely to some limit random variable, we deduce that π_n also converges almost surely and to the same limiting random variable.

Finally, Hardy–Weinberg equilibrium follows readily from the fact that $(\pi_{00}, \pi_{01}, \pi_{11})$ is distributed on the set $\mathcal{S}$, i.e. $F(\pi_{00}, \pi_{01}, \pi_{11}) = 0$.□

Let us now consider the general case $m \geq 1$. We denote by π_G the frequency of a genotype $G = (G_1, \ldots, G_m) \in \{00, 01, 11\}^m$. If $\pi_{i,00}$, $\pi_{i,01}$ and $\pi_{i,11}$, are respectively the limiting gene frequencies of the i-th gene with alleles 0 and 1, then from the assumption of independence between genes (see condition (c) in the previous subsection), the limiting frequency of the genotype G at equilibrium is

$$\pi_G = \pi_{1,G_1}\pi_{2,G_2}\ldots\pi_{m,G_m}.$$

Remark 1. Observe that from equations $\pi_{00} + \pi_{01} + \pi_{11} = 1$ and $\pi_{01} = 2\sqrt{\pi_{00}\pi_{11}}$, the distribution of the limit triplet $(\pi_{00}, \pi_{01}, \pi_{11})$ is actually one dimensional. However, it is a quite challenging question to determine the exact distribution of π_{00}. Our simulations show that it may have a diffuse distribution on [0, 1] which depends on the initial values π_{00}^0, π_{01}^0 and π_{11}^0, see Fig. 4.

Remark 2. A subsequent question to Theorem 1 concerns the speed of convergence of $(\pi_{00}^n, \pi_{01}^n, \pi_{11}^n)$. Some results in this direction are given in Delyon (1996) and Higueras et al. (2006). However, they require some strong assumptions on the derivative of the function F at the limiting point $(\pi_{00}, \pi_{01}, \pi_{11})$, which are quite difficult to verify in our situation, mainly due to the fact that we do not know the distribution of $(\pi_{00}, \pi_{01}, \pi_{11})$. It is reasonable to expect that a central limit type theorem holds, in which case, the speed of convergence of $(\pi_{00}^n, \pi_{01}^n, \pi_{11}^n)$ to $(\pi_{00}, \pi_{01}, \pi_{11})$ would be of order $\sqrt{n}$.

References

Auvray, G., 2011. Les relations phylogénétiques au sein dun systme réticulé: cas particulier de *Cytisus scoparius* L. (Genisteae, Fabaceae) et des espèces, hybrides et cultivars apparentés (Ph.D. thesis), Angers University.

Blouin, M.S., 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol. Evol. 18, 503–511.

Delyon, B., 1996. General results on the convergence of stochastic algorithms. IEEE Trans. Autom. Control 41, 1245–1255.

Geyer, C.J., Ryder, O.A., Chemnick, L.G., Thompson, E.A., 1993. Analysis of relatedness in the California condors from DNA fingerprints. Mol. Biol. Evol. 10, 571–589.

Higueras, I., Moler, J., Plo, F., San Miguel, M., 2006. Central limit theorems for generalized Pólya urn models. J. Appl. Probab. 43 (4), 938–951.

Jeffreys, A.J., Turner, M., Debenham, P., 1991. The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships,

determined from extensive casework. Am. J. Hum. Genet. 48, 824–840.
Jones, B., 2003. Maximum likelihood inference for seed and pollen dispersal distributions. J. Agric. Biol. Environ. Stat. 8, 170–183.
Kirkpatrick, B., Li, S.C., Karp, R.M., Halperin, E., 2011. Pedigree reconstruction using identity by descent. J. Comput. Biol. 18 (11), 1481–1493.
Pemantle, R., 2007. A survey of random processes with reinforcement. Probab. Surv. 4, 1–79.
Pradeep Reddy, M., Sarla, N., Siddiq, E.A., 2002. Inter simple sequence repeat (ISSR) polymorphism and its application in plant breeding. Euphytica 128, 9–17.
Schreiber, S., 2001. Urn models, replicator processes, and random genetic drift. SIAM J. Appl. Math. 61 (6), 2148–2167.
Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford, UK.
Steel, M., Hein, J., 2006. Reconstructing pedigrees: a combinatorial perspective. J. Theor. Biol. 240 (3), 360–367.
Steel, M., Hendy, M.D., Penny, D., 1998. Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results. Discret. Appl. Math. 88, 367–396.
Thatte, B.D., 2013. Reconstructing pedigrees: some identifiability questions for a recombination–mutation model. J. Math. Biol. 66, 1–2.
Thatte, B.D., Steel, M., 2008. Reconstructing pedigrees: a stochastic perspective. J. Theor. Biol. 251 (3), 440–449.
Thompson, E.A., Meagher, T.R., 1987. Parental and sib likelihoods in genealogy reconstruction. Biometrics 43, 585–600.
Thompson, E.A., 1976. Inference of genealogical structure. III. The reconstruction of genealogies. Soc. Sci. Inform. 15, 507–526.
Thompson, E.A., 2000. Statistical inference from genetic data on pedigrees. In: NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 6, Institute of Mathematical Statistics, Beachwood, OH. American Statistical Association, Alexandria, VA.
Wakeley, J., King, L., Low, B.S., Ramachandran, S., 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190, 1433–1445.
Wolfe, A.D., Xiang, Q.-Y., Kephart, S.R., 1998. Assessing hybridization in natural populations of Penstemon (Scrophulariaceae) using hypervariable intersimple sequence repeat (ISSR) bands. Mol. Ecol. 7, 1107–1125.