

Reconstructing pedigrees using probabilistic analysis of ISSR amplification. *

Loïc Chaumont¹, Valéry Malécot², Richard Pymar³ and Chaker Sbai⁴

August 4, 2015

¹ LAREMA – UMR CNRS 6093, Université d’Angers, 2 bd Lavoisier, 49045 Angers Cedex 01

² IRHS – UMR 1345, Agrocampus Ouest Angers, 2 rue Le Nôtre, 49045 Angers Cedex 01

³ Department of Mathematics – University College London, Gower Street, London WC1E 6BT

⁴ PEGASE – UMR 1348, Agrocampus Ouest Rennes, 65 rue de Saint-Brieuc, CS 84215, 35042 Rennes Cedex

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*.

Keywords: Pedigree, ISSR amplification, law of reproduction, gene frequency

Mathematics Subject Classification (2000): 92D25; 92D10; 60F15

*This work was supported by MODEMAVE research project from the Région Pays de la Loire. Acces 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.

1 Introduction

A pedigree is a graph such that each vertex has indegree equal to 0 or 2 and any out-degree. When it represents family relationships between living individuals, edges are directed from parents to children. By reconstruction of the pedigree of a family of some set of individuals, we mean a way to determine the most likely pedigree relating these individuals given some information such as phenotype, genotype, date of birth, data obtained from professional breeders,... It may happen that this information is known only for a part of the population or even that the number of missing individuals is unknown. To each situation corresponds some specific methods. Deterministic methods based on the maximum parsimony principle and using purely combinatorial arguments allow us to reconstruct the minimal pedigree relating individuals in accordance with their types, see Chapter 4 in [10], [11] or [2]. There are also numerous different stochastic methods of reconstruction of pedigrees, see for instance [6], [13], [14], [2]. In any case, the method consists in finding a 'nice' probabilistic framework in which we may find the most likely pedigree relating some set of individuals. Some 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, [6]. 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, [12] and [13]. Coalescence theory may also prove to be a powerful tool in reconstruction of pedigrees, as observed in [15].

In the present work, we assume that the known information is of a genomic type and is provided through ISSR amplification for diploid plant cultivars, which are vegetatively propagated. ISSR amplification was popularised by [16] and largely used in genetic diversity assessment [8]. Because being vegetatively propagated, the available dataset contains both descendants and ancestors in the pedigree, thus both terminal and internal nodes of the graph, while most above listed methods use information from last generation descendants (i.e. terminals in the graph). We know the same genotypic information for each individual and we assume 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 particular, 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). It actually stems as if we observed the phenotypic expression of a dominant gene and our model can also be applied to this kind of situation (see the discussion at the end of this paper). Then from this partial information we provide a probabilistic method to reconstruct the pedigree corresponding to some families of diploid plant 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. More specifically, if g_1, \dots, g_n are individuals ranked in their birth order, then for each $i = 1, \dots, n$, we are looking for a couple of individuals possibly non distinct in the set $\{g_1, \dots, g_{i-1}\}$ which is

the most likely parent pair of g_i 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. It only depends on the gene frequencies which are supposed to be constant in time. In order to justify this assumption, we prove here 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*, *Cytisus* x *boskoopii*). The latter are diploid sexed plants whose crossbreedings have occurred in the past 200 years from a set of founders which is to be specified by our model. 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 as well as to the convergence result of gene frequencies which justifies its relevance. Then 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 represent a pedigree as a directed graph in which each vertex corresponds to an individual and each directed edge corresponds to a parent-child relationship, with the edge going from parent to child. The individuals are partitioned into two sets, F and F^c , referred to as the founders and the non-founders respectively. The pedigree specifies, for every non-founder individual, two (not necessarily distinct) individuals which, according to some probabilistic model shortly defined, are the most likely parents.

We first define the law of reproduction in the population. Let n be the number of individuals, denoted g_1, \dots, g_n and let $m \in \mathbb{N}$ be the number of genes for which we observe the presence or absence of a specific allele. More specifically, when proceeding to the ISSR amplification, for each gene, we receive from some marker, a binary response: either the allele is present in at least one of the two chromosomes or it is absent in both. In particular, when the allele is present, we do not know if it is present on the two chromosomes. Actually, it is equivalent to consider that the allele which is highlighted by the marker is dominant and that we only observe the phenotype of the individual. For each individual g_i and each gene $\ell \in \{1, \dots, m\}$, let $x_\ell(g_i) \in \{0, 1\}$ be the indicator of band absences (0-values) and presences (1-values) of individual g_i obtained during the ISSR amplification

process. Hence the *apparent genotype* of each individual g will be identified to the element $x(g) := (x_1(g), x_2(g), \dots, x_m(g))$ of $\{0, 1\}^m$. 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".

Each individual g has an associated date of birth, denoted $t(g)$. We set $t(g) = 0$ if the individual g was obtained from the wild, in which case it will be considered as a founder. Otherwise set $t(g)$ equal to the date the individual was accessioned. We order the individuals so that for $i < j$, $t(g_i) < t(g_j)$, whenever $t(g_j) > 0$ (it is assumed that dates of birth are distinct from each other). The basic principles of our reconstruction method are:

- (a) a uniform prior on probability (g_j, g_k) are the parents of individual g_i over all pairs (g_j, g_k) with $\max(t(g_j), t(g_k)) < t(g_i)$;
- (b) no missing individuals, that is the parents of each non-founder individual g_i belong to the set $\{g_1, \dots, g_n\} \setminus \{g_i\}$.

Let us denote by \hat{g} and \bar{g} the parents of the individual g . When they breed, the two parents \hat{g} and \bar{g} with respective apparent genotypes $x(\hat{g})$ and $x(\bar{g})$ will give birth to the individual g with apparent genotype $x(g)$ according to the following rules:

- (c) independence of the coordinates of $x(g)$, that is, $\{x_\ell(g) = 1\}$ and $\{x_{\ell'}(g) = 1\}$ are independent for all $\ell' \neq \ell$;
- (d) there are constants $\delta \in (-1/2, 1/2)$ and $\varepsilon \in (0, 1/2)$ called the *errors* and for each ℓ , there are constants $p_\ell \in (3/4, 1)$ and $q_\ell \in (1/2, 1)$ such that for each individual g and

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

Principles (a) and (b) should rather be considered as the most natural assumptions in the absence of any particular constraint in the evolution of the population. Note that according to (a), the father and mother can be the same individual, which is standard in plant populations. Principle (c) means that the evolutions of different genes are independent between each other. In our specific example we will select a particular set of genes whose independence will be checked by means of a statistical test, see Section 3.

Let us now concentrate ourselves on principle (d). Constants δ and ε are actually experimental errors, so they do not depend on gene ℓ . It appears that when the parents

satisfy $\{x_\ell(\hat{g}) = 1\}, \{x_\ell(\bar{g}) = 1\}$ (resp. $\{x_\ell(\hat{g}) = 0\}, \{x_\ell(\bar{g}) = 1\}$), the probability to observe $\{x_\ell(g) = 1\}$ for the child is less than the theoretical probability p_ℓ (resp. q_ℓ), that is $p_\ell - \delta$ (resp. $q_\ell - \delta$). Similarly, it can happen that when the parents satisfy $\{x_\ell(\hat{g}) = 0\}, \{x_\ell(\bar{g}) = 0\}$ one observes $\{x_\ell(g) = 1\}$ for the child. This defines error ε . As showed hereafter, we have $p_\ell \in (3/4, 1)$ and $q_\ell \in (1/2, 1)$, and the estimation from our data, see Section 3, shows that δ and ε are actually of order 0.1.

Besides, we recall that despite the reproduction is sexed, since we are concerned with plant populations, each individual can either be male or female, so that 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 $\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 0\}, \{x_\ell(\bar{g}) = 1\}) = \mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 1\}, \{x_\ell(\bar{g}) = 0\})$.

We now focus on the computation of the conditional probabilities appearing in (d). In order to compute the theoretical values p_ℓ and q_ℓ , let us assume that there is no experimental error, i.e. $\delta = \varepsilon = 0$, so that expressions in (d) are $\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 1\}, \{x_\ell(\bar{g}) = 1\}) = p_\ell$ and $\mathbb{P}(\{x_\ell(g) = 1\} | \{x_\ell(\hat{g}) = 0\}, \{x_\ell(\bar{g}) = 1\}) = q_\ell$. Let us now compute p_ℓ and q_ℓ in terms of the gene frequencies. We will prove in the next section that for each gene, the frequencies of the three genotypes 00, 01 and 11, converge toward some equilibrium, as the number of crossbreeding increases. Let us denote respectively by $\pi_{00}(\ell)$, $\pi_{01}(\ell)$ and $\pi_{11}(\ell)$ these frequencies. Then in our model, we assume that this equilibrium is attained, so that:

(e) $\pi_{00}(\ell)$, $\pi_{01}(\ell)$ and $\pi_{11}(\ell)$ do not depend on time.

Note that here, by time, we mean a scale which is incremented by successive crossbreedings. Assumption (e) will be justified in the next section. When no confusion is possible, we will forget about the index ℓ in $\pi_{00}(\ell)$, $\pi_{01}(\ell)$ and $\pi_{11}(\ell)$. Let us compute p_ℓ and q_ℓ in terms of π_{00} , π_{11} and π_{01} . For a pair of parents (\hat{g}, \bar{g}) chosen uniformly at random in the sub-population $\{g' : t(g') < t(g)\}$, the probability to observe $x_\ell(\hat{g}) = 1$ and $x_\ell(\bar{g}) = 1$ is

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

When they breed and give a child g , the probability to observe $x_\ell(g) = 1$, $x_\ell(\hat{g}) = 1$ and $x_\ell(\bar{g}) = 1$ is

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

We obtain that at any time, p_ℓ is given by

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

Then q_ℓ is obtained in the same way:

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

The frequencies π_{00} , π_{01} and π_{11} belonging 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 $p_\ell = q_\ell(2 - q_\ell)$. In Theorem 1, we show that in fact the triplet of gene frequencies $(\pi_{00}, \pi_{01}, \pi_{11})$ satisfies the Hardy-Weinberg equilibrium, that is $\pi_{01} = 2\sqrt{\pi_{00}\pi_{11}}$ and using this relation, we deduce that

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

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

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

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

$$\begin{aligned} p_{ijk}^{(\ell)} &= \mathbf{1}_{\{x_\ell(g_j)=x_\ell(g_k)=x_\ell(g_i)=1\}}, & \bar{p}_{ijk}^{(\ell)} &= \mathbf{1}_{\{x_\ell(g_j)=x_\ell(g_k)=1, x_\ell(g_i)=0\}}, \\ q_{ijk}^{(\ell)} &= \mathbf{1}_{\{x_\ell(g_j) \neq x_\ell(g_k), x_\ell(g_i)=1\}}, & \bar{q}_{ijk}^{(\ell)} &= \mathbf{1}_{\{x_\ell(g_j) \neq x_\ell(g_k), x_\ell(g_i)=0\}}, \\ \varepsilon_{ijk} &= \sum_{\ell=1}^m \mathbf{1}_{\{x_\ell(g_j)=x_\ell(g_k)=0, x_\ell(g_i)=1\}}, & \bar{\varepsilon}_{ijk} &= \sum_{\ell=1}^m \mathbf{1}_{\{x_\ell(g_j)=x_\ell(g_k)=x_\ell(g_i)=0\}}. \end{aligned}$$

Now define $p_{\delta,\ell} = \min(p_\ell - \delta, 1)$, $q_{\delta,\ell} = \min(q_\ell - \delta, 1)$, $\bar{p}_{\delta,\ell} = 1 - p_{\delta,\ell}$, $\bar{q}_{\delta,\ell} = 1 - q_{\delta,\ell}$, $\bar{\varepsilon} = 1 - \varepsilon$ and

$$\nu_i(j, k) = \begin{cases} \varepsilon^{\varepsilon_{ijk}} \cdot \bar{\varepsilon}^{\bar{\varepsilon}_{ijk}} \prod_{\ell=1}^m p_{\delta,\ell}^{p_{ijk}^{(\ell)}} \cdot \bar{p}_{\delta,\ell}^{\bar{p}_{ijk}^{(\ell)}} \cdot q_{\delta,\ell}^{q_{ijk}^{(\ell)}} \cdot \bar{q}_{\delta,\ell}^{\bar{q}_{ijk}^{(\ell)}}, & \text{if } j \leq k < i, \\ 0, & \text{otherwise.} \end{cases}$$

Then for each $i = 2, \dots, n$, the probability measure μ_i on $\{1, \dots, n\}^2$ is explicitly defined in terms of x by

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

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 $t(g_i) > 0$. Moreover, individuals g_i such that $t(g_i) = 0$ are necessarily founders (i.e. $g_i \in F$), hence their parents do not belong to the current pedigree, so in this case, we set

$$\mu_1(j, k) = 0, \quad j, k \in \{1, \dots, n\}.$$

Fix a *threshold probability* $p \in (0, 1)$. Then an individual g_i is in the set F^c of non founder individuals, only if there exists a pair $(j, k) \in \{1, \dots, n\}^2$ such that $\mu_i(j, k) \geq p$ with

$j \leq k < i$ (it follows that the partitioning depends on the value of p).

For each individual $g_i \in F^{\mathcal{G}}$, we wish to determine g_j and g_k (possibly equal), such that the following two conditions are satisfied:

1. $j \leq k < i$ (g_j and g_k accessioned before g_i);
2. $\mu_i(j, k) = \max_{j', k'} \{\mu_i(j', k') : j' \leq k' < i\}$ (g_j and g_k maximize the likelihood).

We remark that by definition of $F^{\mathcal{G}}$, it follows that if we have found such a pair g_j and g_k , then $\mu_i(j, k) \geq p$.

Note also that the normalization of the probability measure μ is relevant only for the comparison with the threshold probability. Steps 1. and 2. define the algorithm from which we performed the program in R which provides the reconstructions of pedigrees, see Section 3.

2.2 Convergence to equilibrium

In this subsection, we are interested in the dynamics of the frequencies of each genotype in the population. As already mentioned in the previous section, our reconstruction method strongly bears on the assumption that the frequencies π_{00} , π_{01} and π_{11} of the types 00, 01 and 11 do not depend on time, that is condition (e) in subsection 2.1. We will show in the present subsection that as the number of crossbreeding goes on, these frequencies converge almost surely to some random equilibrium. This result actually justifies assumption (e).

From time $n = 0$, we rank the crossbreedings in increasing order as they occur. Since the evolutions of genes are independent of each other, see assumption (c), we only need to consider the dynamics of the frequencies of genotypes 00, 01, 11 for one gene. Then let us denote by π_{00}^n , π_{01}^n and π_{11}^n , the proportion of individuals g with genotype 00, 01 or 11 respectively, after the n -th crossbreeding. Let us 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. That assumes in particular that there is no death. Moreover we assume that both alleles exist in the two founders. Then our reproduction law described in (a)-(d) of the previous subsection may actually be represented as a generalized urn model in which the probability of replacement depends on the proportion of individuals in the population, see [7] and the references therein. More specifically, at each step n (crossbreeding), condition (a) tells us that we choose two individuals uniformly at random in the population.

Let us define the polynomial function $F : \{(x, y, z) \in [0, 1]^3 : x + y + z = 1\} \rightarrow \mathbb{R}^3$ 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 π^n 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:

$$\left\{ \begin{array}{l} \pi_{00}^{n+1} = \frac{(n+2)\pi_{00}^n + 1}{n+3} \\ \pi_{01}^{n+1} = \frac{(n+2)\pi_{01}^n}{n+3}, \\ \pi_{11}^{n+1} = \frac{(n+2)\pi_{11}^n}{n+3} \end{array} \right. , \quad \text{with probability } \pi_{00}^n \pi_{01}^n + (\pi_{00}^n)^2 + (\pi_{01}^n)^2 / 4 = F_1(\pi^n),$$

$$\left\{ \begin{array}{l} \pi_{00}^{n+1} = \frac{(n+2)\pi_{00}^n}{n+3} \\ \pi_{01}^{n+1} = \frac{(n+2)\pi_{01}^n + 1}{n+3}, \\ \pi_{11}^{n+1} = \frac{(n+2)\pi_{11}^n}{n+3} \end{array} \right. , \quad \text{with probability } \pi_{00}^n \pi_{01}^n + \pi_{01}^n \pi_{11}^n + 2\pi_{00}^n \pi_{11}^n + (\pi_{01}^n)^2 / 2 = F_2(\pi^n),$$

$$\left\{ \begin{array}{l} \pi_{00}^{n+1} = \frac{(n+2)\pi_{00}^n}{n+3} \\ \pi_{01}^{n+1} = \frac{(n+2)\pi_{01}^n}{n+3}, \\ \pi_{11}^{n+1} = \frac{(n+2)\pi_{11}^n + 1}{n+3} \end{array} \right. , \quad \text{with probability } \pi_{01}^n \pi_{11}^n + (\pi_{11}^n)^2 + (\pi_{01}^n)^2 / 4 = F_3(\pi^n).$$

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 $\pi^0 = i_0, \dots, \pi^n = i_n$ is $F(i_n)$. Recall that the quantity $(n+2)\pi^n$ represents the population size 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

$$\pi^{n+1} = \pi^n + \frac{1}{n+3}(F(\pi^n) - \pi^n + \eta_n). \quad (2.2)$$

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

$$\begin{cases} \frac{d}{dt} f_u(t) = F(f_u(t)), & t \geq 0, \\ f_u(0) = u. \end{cases} \quad (2.3)$$

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

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

We aim to show almost-sure convergence of $\pi^n = (\pi_{00}^n, \pi_{01}^n, \pi_{11}^n)$ as $n \rightarrow \infty$. The first step in achieving this is to show almost-sure convergence of $v(\pi^n)$ as $n \rightarrow \infty$, where $v(u) := \lim_{t \rightarrow \infty} f_u(t)$. This is achieved in the following lemma.

Lemma 1. *As $n \rightarrow \infty$, $v(\pi^n)$ converges almost surely.*

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

$$|v(\pi^{n+1}) - v(\pi^n)| \leq \left| v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right) - v(\pi^n) \right| + \left| v(\pi^{n+1}) - v\left(\pi^n + \frac{1}{n+3}F(\pi^n)\right) \right|. \quad (2.4)$$

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

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

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 (2.4), 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 (2.2). However since F is bounded we deduce that we can upper bound this term by $O(1/n)$. Plugging the two bounds we have obtained into equation (2.4) shows that the sequence $(v(\pi^n))_n$ is indeed Cauchy (surely), and this completes the proof. \square

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:*

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

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 $n \rightarrow \infty$. 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 [9] 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 (2.3). 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 $x \rightarrow \mathcal{S}$, $|v(x) - x| \rightarrow 0$. 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. $F(\pi_{00}, \pi_{01}, \pi_{11}) = 0$. \square

In this theorem, an additional information is brought by the Hardy-Weinberg principle which provides a relationship between the allelic frequencies and the genotypic frequencies. This equilibrium was predictable and is actually a natural consequence of the absence of any evolutive forces.

Let us now consider the general case $m \geq 1$. We denote by π_G the frequency of a genotype $G = (G_1, \dots, G_m) \in \{00, 01, 11\}^m$. 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 independence between genes (see condition (c) in the previous subsection), the limiting frequency of the genotype G at equilibrium is

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

Remark 1. *It is a quite challenging question to determine the exact distribution of the limit triplet $(\pi_{00}, \pi_{01}, \pi_{11})$. Actually our simulations show that it may have a diffuse distribution in the set $\{(x, y, z) \in [0, 1]^3 : x + y + z = 1\}$, which depends on the initial values π_{00}^0 , π_{01}^0 and π_{11}^0 , see Figure 1.*

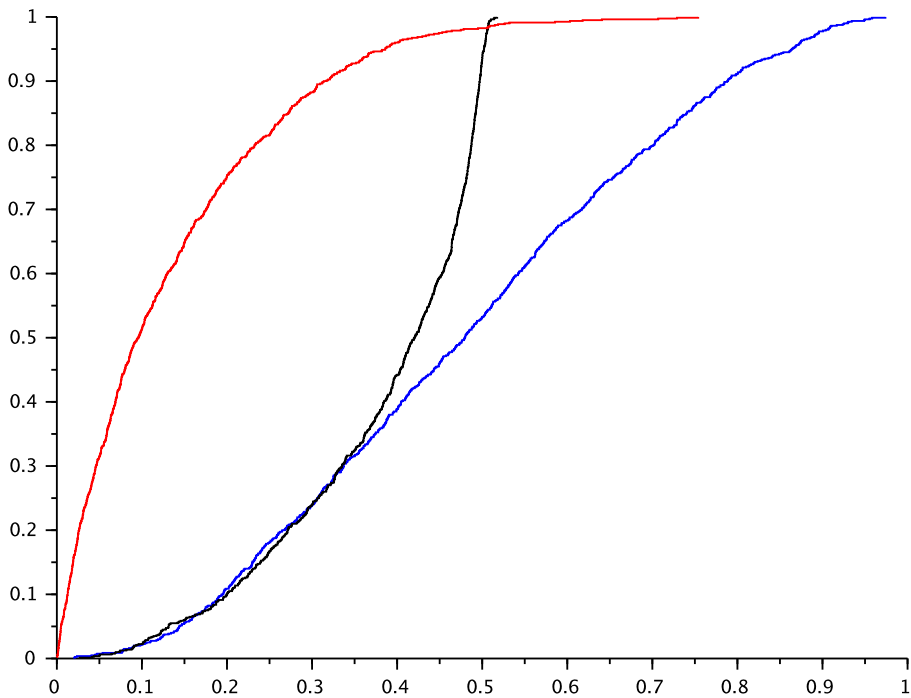
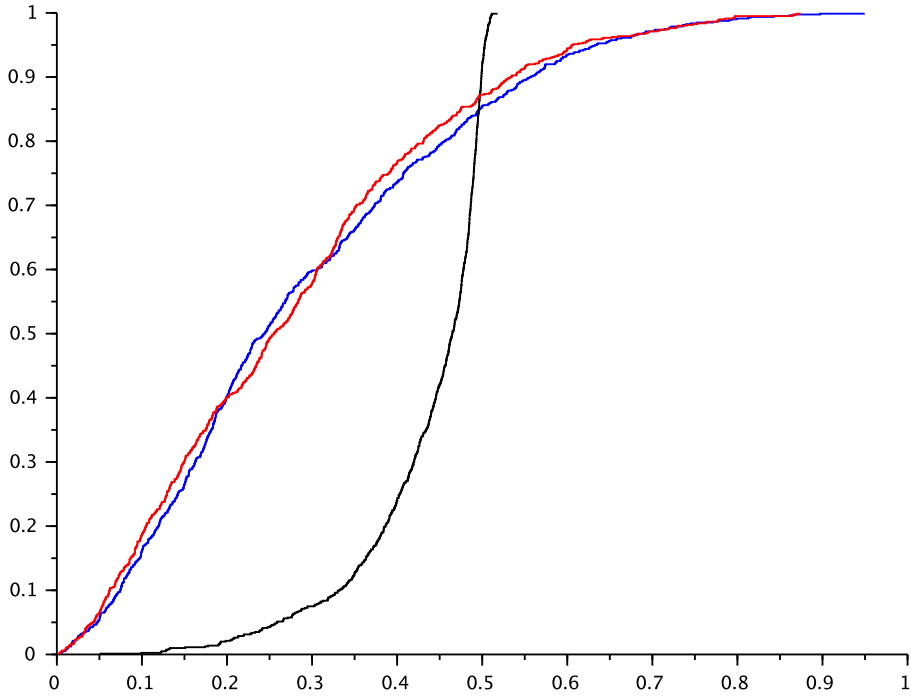


Figure 1: Empirical distribution functions of π_{00} (blue), π_{11} (red) and π_{01} (black). The first figure is obtained with initial values $\pi_{00}^0 = 1$, $\pi_{01}^0 = 2$, $\pi_{11}^0 = 3$ and the second one is obtained with $\pi_{00}^0 = 1$, $\pi_{01}^0 = 1$, $\pi_{11}^0 = 0$.

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 [3] and [4]. 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 $(\pi_{00}, \pi_{01}, \pi_{11})$. However, 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} .

3 Application of the model

Our model were tested on a population of 85 living accessions representing the common broom *Cytisus scoparius* and three related interspecific hybrids. This dataset consists in 62 vegetatively propagated cultivars obtained from various nurseries. These cultivars belong to either *Cytisus 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 five wild populations have been included (3 individuals of *Cytisus oromediterraneus* from France, 3 individuals of *Cytisus scoparius* from Italia, 3 from Poland, 4 from Angers, France and 9 from Ernée, France). For all these samples, DNA extration use the Nucleospin®Plant II kit from macherey-Nagel. IISR data was obtained using six set of primers, namely ISSR5 (sequence: 5-CACACACACACACARC-3), ISSR7 (sequence : 5-CACACACACACACART-3), ISSR13 (sequence: 5-GTGTGTGTGTGTGTGTGTYA-3), ISSR890 (sequence: 5-VHVGTGTGTGTGTGTGTGT-3), ISSR891 (sequence : 5-HVHTGTGTGTGTGTGTGTG-3) and ISSRa (sequence: 5-GCTCTCTCTCTCTC-3). Polymerase chain reaction (PCR) was done using the following parameters : 95°C for 2 min., then 39 cycles of 95°C for 30 sec., 50°C for 30 sec., 72°C for 120 sec., followed by 10 min. of extension at 72°C. Electrophoresis was done on 5% acrylamide-bisacrylamide gel (mixing ratio : 29:1), with 7M urea, with a pre-run of 30 min at 80 W, then 2h30 at 60W. Staining use silver nitrate. Gels were scanned and band manually read.

Using data obtained from ISSR analysis, our present aim 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. The latter applied to our data provided the pedigrees presented in figures 2, 3 and 4 below. The use of this method first requires that the population we are dealing with satisfies principles (a) – (e) in Subsection 2.1 and parameters ε , δ , p_ℓ and q_ℓ must be inferred from our data.

Breedings have occurred over time under the action of professional breeders or according to natural phenomenons and with no more information, assumption (a) about uniform prior distribution is reasonable. According to botanists, this is also the case of assumption (b) which means that there are no missing individuals in the population. Then we need to ensure the independence hypothesis (c) between the bands $\{x_\ell(g) = 1\}$, $\ell \in \{1, \dots, m\}$.

Depence 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. Then among the 424 bands, we have selected 168 of them which are proved to be independent from a statistical test.

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 which is defined in (d). First recall that in the ISSR amplification, six markers allow us to test the presence or absence of those 168 bands, each marker corresponding to a particular set of bands (34 bands for ISSR890, 22 for ISSR 891, 31 for ISSRa, 32 for ISSR5, 27 for ISSR7 and 22 for ISSR13). For each of the six markers used, in order to apply the above model, we need to estimate the values of δ and ε (the errors probability, which can occur during the experiment). We achieve this by repeatedly crossing two individuals (G017 *Cytisus scoparius* 'Lunagold' and G010 *Cytisus x dallimorei* 'Burkwoodii') and performing marker analysis (using 5 of the 6 markers used for the dataset) on the resulting offspring (n=33 plants). We are then able to estimate, for each marker, the value of δ . Denoting by δ_m the error using marker m , we assume that δ_m is a Gaussian random variable such that $\text{Var}(\delta_m) = \text{Var}(\delta_{m'})$ for all markers m, m' . We obtained the following average errors:

$$\begin{aligned}\mathbb{E}(\delta_{ISSRa}) &= 0.16, \mathbb{E}(\delta_{ISSR890}) = 0.16, \\ \mathbb{E}(\delta_{ISSR891}) &= 0.14, \mathbb{E}(\delta_{ISSR5}) = 0.19, \mathbb{E}(\delta_{ISSR7}) = 0.1.\end{aligned}$$

For each pair of markers, m and m' , we ran a hypothesis test to determine whether $\mathbb{E}(\delta_m) = \mathbb{E}(\delta_{m'})$ and we found that we do not reject this null hypothesis at a 95% confidence level. We obtained a 95% confidence interval of (0.126, 0.195) for the error, under the assumption that the errors from the different markers all came from the same distribution. For the present reconstructions we have chosen the value $\delta = 0.15$. The same study for the error ε leads us to the choice of $\varepsilon = 0.05$.

In subsection 2.2 we proved convergence of gene frequencies and we will assume that the population which is considered here has attained some equilibrium, that is principle (e). As can be seen from equation (2.1), thanks to Hardy-Weinberg principle, the probabilities p_ℓ and q_ℓ only depend on the probability π_{00} . We emphasize 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 Subsection 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 figures 2, 3 and 4 were obtained with the threshold probabilities 0.1 and 0.2 and 0.3 respectively. Funders have been represented in black and individuals with no parent and children have

not been represented. 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 [1]), 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 ardoinei* 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-fecundation 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 [1] 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 [1] 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 Favorite' 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 Favorite' and 'Queen Mary', appearing under all threshold probabilities, reminds 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 the probability distribution which is defined in (d) of Subsection 2.1. The robustness of this model mainly relies on the fact that gene frequencies have attained some equilibrium. We show in Subsection 2.2 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

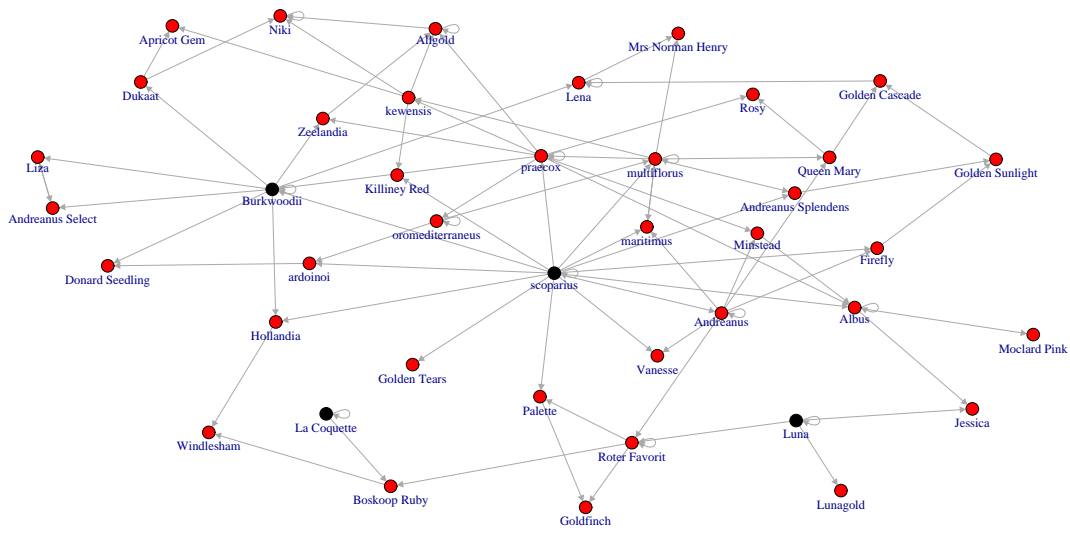


Figure 2: Threshold probability $p = 0.1$.

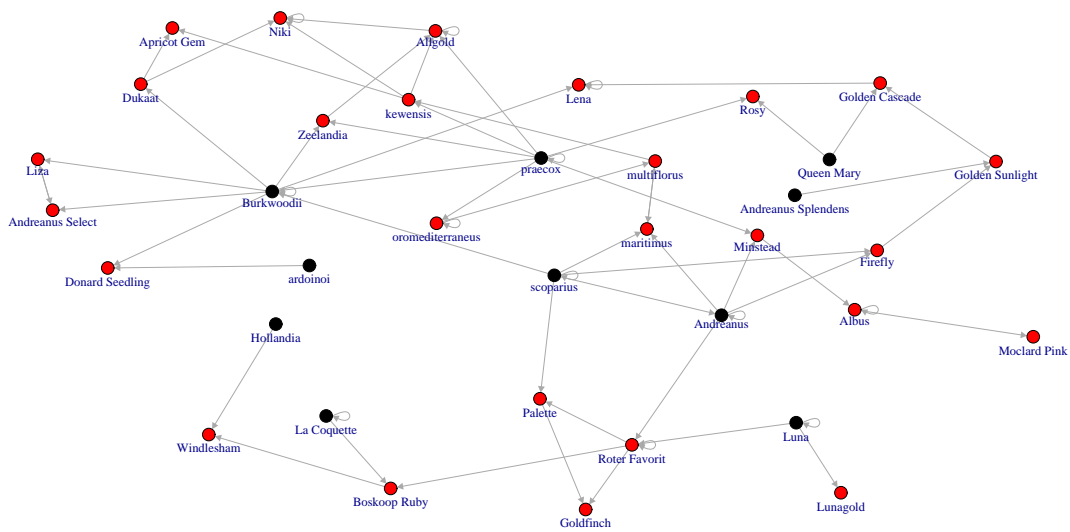


Figure 3: Threshold probability $p = 0.2$.

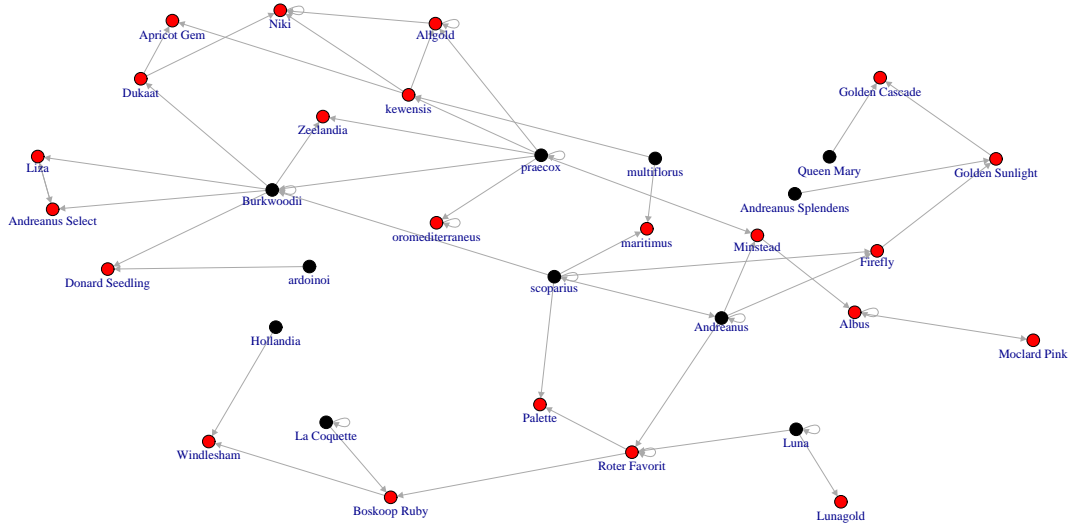


Figure 4: Threshold probability $p = 0.3$.

reveal that the pedigrees obtained from this method fit to the partial reconstructions based on botanical data or other methods using dendrograms 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.

Greater power could also be given to our method by getting rid of assumption (b) on non missing individuals. Indeed missing individuals in the population who would actually have lots of family relationships could considerably distort the real pedigree. Then an improvement would consist in determining how much the addition of one or several virtual individuals with specific genomes increases the likelihood of the pedigree.

Principle (c) assumes that recombination is uniform, but this can be made more realistic by determining how different sets of loci actually recombines from a preliminary statistical inference. Then the model can easily be adapted.

Finally we emphasize that our model can be applied to phenotyped 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.

References

- [1] G. AUVRAY. Les relations phylogénétiques au sein dun systme réticulé : cas particulier de *Cytisus scoparius* L. (Genisteae, Fabaceae) et des espèces, hybrides et cultivars apparentés. *PhD Thesis*, Angers University, (2011).
- [2] M.S. BLOUIN: DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution* 18: 503-511, (2003).
- [3] B. DELYON: General results on the convergence of stochastic algorithms. *IEEE Trans. Automatic Control*, 41:1245–1255, (1996).
- [4] I. HIGUERAS, J. MOLER, F. PLO AND M. SAN MIGUEL: Central limit theorems for generalized Pólya urn models. *J. Appl. Probab.*, **43**, no. 4, 938–951, (2006).
- [5] B.M. HILL, D. LANE AND W. SUDDERTH: A strong law for some generalized urn processes. *Ann. Probab.*, **8**, no. 2, 214–226, (1980).
- [6] B. KIRKPATRICK, S.C. LI, R.M. KARP AND E. HALPERIN: Pedigree reconstruction using identity by descent. *J. Comput. Biol.*, **18**, no. 11, 1481–1493, (2011).
- [7] R. PEMANTLE: A survey of random processes with reinforcement. *Probab. Surv.*, 4, (2007), 1–79.
- [8] M. PRADEEP REDDY, N. SARLA, E.A. SIDDIQ: Inter simple sequence repeat (ISSR) polymorphism and its application in plant breeding. *Euphytica*, 128 : 9–17, (2002).
- [9] S. SCHREIBER: Urn models, replicator processes, and random genetic drift. *SIAM J. Appl. Math.* 61, no. 6, 2148–2167, (2001).
- [10] C. SEMPLE AND M. STEEL: *Phylogenetics*. Oxford University Press, 2003.
- [11] M. STEEL AND J. HEIN: Reconstructing pedigrees: a combinatorial perspective. *J. Theoret. Biol.*, **240**, no. 3, 360–367, (2006).
- [12] M. STEEL, M.D. HENDY AND D. PENNY: Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results. *Discrete Applied Mathematics*, **88**, 367–396, (1998).
- [13] E. A. THOMPSON: Statistical inference from genetic data on pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics, 6. *Institute of Mathematical Statistics, Beachwood, OH; American Statistical Association, Alexandria, VA*, 2000.
- [14] B.D. THATTE AND M. STEEL: Reconstructing pedigrees: a stochastic perspective. *J. Theoret. Biol.*, **251**, no. 3, 440–449, (2008).

- [15] J. WAKELEY, L. KING AND B.S. LOW, AND S. RAMACHANDRAN. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics* 190:1433-1445, (2012).
- [16] A.D. WOLFE, Q-Y. XIANG, S.R. KEPHART: Assessing hybridization in natural populations of *Penstemon* (Scrophulariaceae) using hypervariable intersimple sequence repeat (ISSR) bands. *Molecular Ecology*, 7 : 1107–1125, (1998).