Markovian simulation for ancestors trees

C. Jarne
Departamento de Ciencia y Tecnología Universidad Nacional de Quilmes (UNQ) and IFIBA-CONICET, 1428 Buenos Aires, Argentina.

M. Caruso
Departamento de Física Teórica y del Cosmos, Universidad de Granada, Campus de Fuentenueva, Spain

(Dated: March 10, 2017)

We present a computational model to reconstruct trees of ancestors for animals with sexual reproduction following the theoretical model presented in Phys. Rev. E 90, 022125 (2014). Through a recursive algorithm combined with a random number generator, it is possible to reproduce the number of ancestors for each generation and use it to constraint the maximum number of the following generation. This simulation allows to consider the reproductive preferences of particular species and combine several trees to simulate the population behavior. It is also possible to describe the simulation analytically, as a theoretical stochastic process. The obtained process can be generalized in order to use the algorithm associated with it to simulate similar processes of stochastic nature.

I. INTRODUCTION

Previous works have shown that it is not possible to reconstruct the genealogical history of each species along all its generations through a geometric progression [1]. Because the geometric progression is determined by a sequence of independent events. From the model point of view, we postulate that blood relationship between ancestors (or inbreeding) is the key to connect the consecutive generations.

It is well known that close inbreeding within species can result in inbreeding depression caused by an increase in homozygosity of recessive, deleterious alleles and loss of heterosis [2]. However, extreme intraspecific outbreeding can also be disadvantageous (outbreeding depression) [3], because beneficial gene complexes or local genetic adaptations are disrupted [4]. There is numerous empirical support that animals avoid close kin as mating partner [5, 6]. Optimal outbreeding theory [8] is supported by experimental behavioral studies showing mating preferences for intermediately related individuals [9] and by genetic studies reporting stabilizing selection on genomic divergence in wild populations of animals [10] and plants [11].

Empirical studies of inbreeding strategy have reported inbreeding tolerance in different species such as New Zealand robins, Petroica australis [12]; bighorn sheep (Ovis canadensis) [13]; great tits (Parus major) [14] and even inbreeding preference in cichlid fish (Pelvicachromis taeniatus) [15]. Besides recent studies have found evidence of regular incest in wild mammals, even in social species where relatives are spatio-temporally clustered and opportunities for inbreeding frequently arise [16].

Considering some degree of incest along biparental species as a key to build a general model and using a Markov process it is possible to obtain a more realistic model for a tree of ancestors [17].

If we consider a random variable that represents the number of ancestors present in a given generation, the size of the sample space depends on each generation. This brings huge problems to find an analytical solution.

In [17] we proposed a method to solve this problem. Now we present an alternative way to obtain an estimation of the mean value of the number of ancestors for each generation following the same time-line approach that [17], meaning to start from the present and track back the possible sequence of ancestors. We used a simple computational model to follow back the chain of ancestors.

Our first model consisted in a simple algorithm that we will describe in the following sections. Before starting the description, there are two important assumptions about the biology of the considered system. Initially, the species described here has no specific behavior of sexual partner selection (random mating reproduction) [18, 19]. This means that in our model the partner could be a kin or not. Many species or population groups exhibit this kind of reproduction. The distribution of ancestors for a given generation is contained in a population large enough to avoid forcing the selection of blood-related sexual partners. The partners could be blood related or not at random, up to a maximum generation when the process start to decrease. Current model presents a random mating in non-overlapping generations with negligible mutation and selection. These two assumptions are common to develop population genetic models; in particular, these are
present in the Hardy-Weinberg principle\cite{20, 21}.

We present a description of the problem through an algorithmic and analytic approach. The structure of this manuscript is developed as follows: Section II is a description of the algorithm implemented to generate the trees of ancestors. Section III shows an equivalent description and a generalization in terms of mathematical symbols. A comparative results between the implemented algorithm and the theoretical model \cite{17} are presented in Section IV. Finally, in Section V we present the conclusions and further work.

II. BUILDING THE COMPUTATIONAL MODEL

II.1. A simplified tree of ancestors

We started creating a simplified tree of ancestors for one individual in the following way. In this work we considered an index $n$ that labeled the generation number, starting from $n = 0$ the parents generation, $n = 1$ the grandparents generation and so forth. Initially we fixed the number of ancestors for the first generation, $n = 0$, at 2, because the individuals have sexual reproduction. In successive generations $n \geq 1$ we run a random number generator, denoted by $r_n$, to obtain a number between 2 and a maximum value that will depend on generation number. In each generation the maximum possible number of ancestors is:

$$R_n = 2^{n+1}. \quad (1)$$

To avoid a high endogamy degree at the beginning of the tree, we also fix this value at generation $n = 1$ at 4. Therefore in the next generation, $n = 2$, the maximum possible ancestor number is $R_2 = 8$. Now at $n = 2$ for the first time we run a random generator with a uniform distribution to obtain a number between 2 and $R_2$. The possible maximum number for ancestors in the following generation, $n = 3$ is less or equal to $R_3$. Actually this value depends on the random value obtained in the generation before.

In general, for one individual at the generation $n$ is not possible to have more ancestors that the double of the ancestors obtained in $n - 1$ (two parents per each predecessor), i.e:

$$r_n \leq 2r_{n-1}. \quad (2)$$

and of course we have $r_n \leq R_n$, for all $n$. The expression \cite{2} is a better upper bound that $r_n \leq R_n$, because $2r_{n-1} \leq R_n$.

To illustrate a set of possible trees, the Figure shows three simple examples for the first generations in the chain.

For each $n -$ generation in the tree, we obtain a random number $r_n$ between the limits given for the random number obtained in the previous generation, i.e. $r_n \in [2, 2r_{n-1}]$, which represents the possibility of blood relationship, or endogamy, between the individuals of the same generation. A full tree is developed via a generation number loop increasing program. In order to avoid any possible bias in the tree we change the seed of the random generator in each run. Additionally we establish a given generation in which the number of ancestors is maximum (more detail presented in Section II.3).

The length of the genealogical tree is as long as we prefer to define, for instance in our illustrative example we use $N = 50$ generations at most, depending on when the random generator obtains a minimum of 2 ancestors. Nevertheless an important consideration is that the simulation of one particular tree could end early if for a given generation the number of ancestors reaches, by chance, the number 2. Four examples of random trees generated are shown in Figure where $R_n$ in \cite{1} corresponds to the case in which all ancestors are different in each $n -$ generation.
With this first approach presented, very endogamous trees has been occasionally obtained. The explanation for this is that any possible combination of inbreeding between ancestors has the same probability (uniform distribution) and becomes very unlikely a non-inbreeding case, therefore we need to introduce additional considerations over the endogamy degree. Therefore we improve the algorithm in order to provide a more realistic simulation.

II.2. A more sophisticated tree of ancestors

We found a way to have more control of the endogamy degree of the tree. On one hand, the algorithm that we used has been enriched by endowing it with the possibility of switch the reproduction between inbreeding or not. A second random number generator was used in each generation, namely \( s_n \), that returns an integer \( s_n = 0 \) or \( s_n = 1 \). The outcome of \( s_n \) and \( r_n \) can affect to \( r_{n+1} \) stochastically. One of the leading actors in this algorithm is the distribution used for the random generator number \( r_{n+1} \), taking into account the previous result for \( r_n \) and the branch from which it comes. Analytically this is equivalent to a conditional probability for \( r_{n+1} \) given \( r_n \) and \( s_n \): \( P(r_{n+1}|r_n, s_n) \). A schematic diagram of the above description is shown in Figure 3.

On the other hand, in this work we considered equal probabilities for each branch, i.e. \( s_n = 0 \) or \( s_n = 1 \): \( P(s_n = 0) = P(s_n = 1) \). Then of course, we have 50% chances to obtain \( s_n = 0 \), also the same chances for \( s_n = 1 \), for each \( n \).

Different reproductive behavior such as one male mating with several females in a group, could be changed via the percent rate in this second generator. Other behavior like sibling selection could be included via the probability distribution of animal preferences using another distribution of probability different of the uniform, to give different weights within the inbreeding selection option, between 2 and \( 2r_n - 1 \).

We explored different distributions of random numbers for the development of each individual tree, but we selected those that could be linked experimentally with an animal reproductive behavior. In this case we study a uniform distribution and a negative exponential distribution for the growth of the tree and a negative exponential distribution for the decay of the tree. These and other more details will be exposed in the following section.

II.3. Details on the tree simulation parameters

Due to the freedom in the algorithm design as well as in its parameters, it is necessary to specify the information that lead to one possible development of the tree. Taking into account the experimental data, most of our assumptions could be replaced by constraints of animal nature. Our intention is that others could be able to use the algorithm, i.e. choose the parameters and the probability distribution, in a way that would be useful to represent real data sets. In this section we present our selection of parameters.

Each simulated tree has a maximum length of \( N \) generations, but of course a given execution may finish before, as we explain in Section II.1. The simulation process is mainly separated into two regimens: growth and decay; where the number of ancestors increases and decreases, on average, respectively. As we explained above in Section II.1 we chose a generation to be the one where the number of ancestors is maximum. We do not chose the maximum number of ancestors, we only chose where this maximum is reached, denoted by \( N_\mu \). After this generation the number of ancestor could never exceed the random value obtained for \( N_\mu \), generation. This generation define the growth interval: \([0, N_\mu] \). In this paper we used \( N_\mu = N/2 \). The decay interval is defined from \( N_\mu \) to \( N \). Both parts of the tree development, growth and decay, are ruled almost by the same algorithm rule: using a second random number generator \( s_n \) to make a bifurcation in the algorithm. All these bifurcations lead to different actions depending on the regime we are dealing with, i.e. growth or decay.

For the growth regime, remember the interval \([0, N_\mu] \), if \( s_n = 0 \), the number of ancestors at generation \( n+1 \) will be the maximum possible with no inbreeding: \( r_{n+1} = 2r_n \). This means that each ancestor in the generation \( n \) has all different parents in the generation \( n+1 \), i.e. the pairs of parents of each individual of generation \( n+1 \) are different one to one. If \( s_n = 1 \), we will have for \( r_{n+1} \) any kind of inbreeding reflected between 2 and \( 2r_n - 1 \). For this branch we use a uniform or a negative exponential distribution for the random number \( r_{n+1} \).

The decay regime is subdivided in two parts, the first one is defined up to a cut generation \( N_\gamma \). In this interval, \([N_\mu, N_\gamma] \), the branches \( s_n = 0 \) and \( s_n = 1 \) are equally the same as in the growth regime, the only difference here is that a negative exponential distribution is used for the random number \( r_{n+1} \). The second one is a more harder decay version of the first part. In the interval \([N_\gamma, N] \) and branch \( s_n = 0 \), the not endogamy rule is replaced by a full-endogamy rule. We define full-endogamy as the case where the number of ancestors is fixed and constraint to the number of individuals at the beginning of the ancestors tree. In an extreme case, the minimal number of individuals prefixed would be 2 or 4, but could be any number. This pre-fixed number represents the upper bound of the initial number of individuals at the beginning of the population who originated the tree. In other
FIG. 2. (Color online). Example of four kinds of genealogical trees from present time to past. Red dots are the number of ancestors for each \( n \)–generation at most equal to \( N = 50 \). All examples are compared with no-inbreeding tree, \( R_n = 2^{n+1} \), represented with the black line.

words, we merge random endogamy probability with the probability of having a low fix number of ancestors; this, again, represent the upper bound of individuals at the beginning of an initial population. The \( s_n = 1 \) branch remains the same, only use a negative exponential distribution for the random number \( r_{n+1} \).

For this reason, we call the first part as \textit{soft decay} and the second part as \textit{hard decay}.

In the Figures presented in this work, the \( N_\gamma \) is selected 10 generations before \( N \). Even when some trees could end by chance before this generation, we used this rule to conduct any surviving tree to converge to the maximum possible value of individual who originated the tree, at \( N \)–generation could be 2, 4 or any other value. The number of ancestors at the end of tree is constrained to the number of individuals who originated a particular population.

To summarize, the development of tree of ancestors could be split in three steps delimited by this four generations ordered as follows: \( 0 < N_\mu < N_\gamma < N \). The \textit{growth} region starting from generation 0 to \( N_\mu \) and for the \textit{decay} region consisting in two intervals: \textit{soft decay} from \( N_\mu \) to \( N_\gamma \) and \textit{hard decay} from \( N_\gamma \) to \( N \).

As we said before, we could add on the simulation a condition to request that not all possible inbreeding options be equally likely. To do that different distribution are used in the growth to represent differences in animal mate preferences. We compared a simulation using the uniform distribution, i.e. no preference in mate selection,
with a simulation using the negative exponential distribution (siblings preference). In the case of the second option, there is an additional free parameter in the distribution that allows us to control the endogamy degree.

The exponential decay case corresponds to a very inbred preference of the ancestors (close kin), where an increasing exponential distribution corresponds to avoid the incest as long as possible (within the random case). The opposite case is an increasing exponential distributions, it corresponds to a preference selection of a not close kin when selecting a kin for mate.

We chose a shift in the negative exponential distribution, that depends on the generation. This shift was chosen in order to obtain the maximum endogamous probability at the half of the interval between 2 and 2\(r_{n-1} - 1\) for each generation \(n\). In the decay part, our selection was to use always the exponential decay distribution, since we want to give in the decay more weight to the endogamous preference.

Other additional distribution could be used, such a gaussian distribution. In this case there is an additional parameter to fix, that does not necessary represent an aspect in animal mate selection. We did not use such distribution because we want that all parameters represent some aspect on the animal mate preference. But even so, the gaussian distribution could be used in other similar simulations based on the ancestors simulation algorithm to represent some other markovian process.

In Section II.3 we show a way to describe the simulation as a theoretical stochastic process, that allows to describe the expected value of \(r_{n+1}\) given the values of \((r_n,s_n)\), as a recurrence equation, using the Law of total probability. The expected value for \(r_{n+1}\) depends on two terms corresponding to the cases \(s_n = 0\) or \(s_n = 1\).

II.4. Building an ensemble of trees

Once we generated the tree for one individual, we studied what happens with a general population with independent ancestors trees generated as explain in Section II.3. The process consists in generating the reproductive history of each individual for a certain number of generations (to create a set of individual random trees) and then take the average of the total population, generation by generation. In that way we have a representative mean behavior of the trees population for each generation.

The average of the generated trees samples could be compared with the expected value of the random variable in the first theoretical model [17].

Our first trial was made considering a population of 50 individuals each one with its own tree, to have an animal population large enough to perform the statistical analysis but at the same time fast to reduce computational time. In this way we can study different effects on the selection of model parameters, i.e. we averaged each generation of the independent 50 trees.

There are other ways to combine the trees without using the mean of the sample set, for instance by means of a genetic algorithm [24]. But the arbitrary choice of a parameter over others to optimize the trees (fitness) and use them as seeds to generate new trees prevents us to follow this approach. Additional studies with data from the biology field are necessary to follow the genetic algorithm approach. Comments on the proposed model from biologist are specially welcome to enlighten us to follow this possible path.

II.5. Results of trees ensembles

In order to perform the simulation we used different distributions for the random number generator as described in Section II.3. The simplest case consist in taking the uniform distribution. The results obtained with uniform distribution are presented in Figure 4 (right). The uncertainties in all Figures are the bin errors associated with each \(y_b\), the data of the \(y - axis\) in the \(b\)-bin. This quantity is calculated as the standard error on the mean as \(\Delta y_b = S(y_b)/\sqrt{n_b}\), where \(S(y_b)\) is the standard deviation (RMS) of the \(y_b\) data and \(n_b\) is the number of bin entries. In this case the entries are the number of trees that are considered in the calculation of the mean at the \(b\)-bin.

Different selection for the generation where the number of ancestors starts to decrease and a maximum generation to end the process, \(N\), previously described in II.3 are shown in Figure 4. In each case, the mean number of ancestors grows, reaches a maximum value and then decreases, depending on the maximum generation to end the process. Figures show process ending up to generation \(N = 20\), increasing this value up to \(N = 30\). When the maximum generation is changed, the evolution of the mean value is affected as it is shown in Figure 4.

Other possible setting for the tree of ancestors is not to finish the process when the number of ancestors is two but also to end it with a different number of ancestors as initial population as we said in Section II.3. This case corresponds to a particular population starting not with 2 original ancestors, but with other number of couples of animals, for instance 200 original individuals as the initial population. This case is shown in Figure 5. This is a parameter that can be changed in order to compare with different animal population under study.

We used a negative exponential distribution, with the same constraints representing a preference for the animals to chose close kins and compare it with the uniform
FIG. 4. Mean number of ancestors for each generation. Average of 50 trees. For each tree the random generator uses an uniform distribution of random numbers. From left to right and from top to bottom the maximum generation in each case is \(N = 20, N = 22, N = 24, N = 26, N = 28\) and \(N = 30\).

case. We expected to obtain smaller values for the number of ancestors in this case.

A comparison between the case using the uniform distribution and the exponential decade distribution are shown in Figure 5, where the average of 50 trees generated with each one are shown with dots and squares. As we expected smaller values for the number of ancestors are obtained when using the exponential distribution for the same values of initial population and maximum generation.

The Figure (Top) corresponds to the case where the initial number of ancestors is different to 2, using the uniform distribution. The Figure (Bottom) corresponds to the comparison between the uniform distribution and the
negative exponential distribution.

\[ \sum_r T(n|r) \frac{T(n)}{T(n)} = 1. \]  

If we have from the simulation a quantity *almost* equal to the \( r_n \) distribution, we can obtain the expected value, or even more, the higher \( k \)—order moments for each generation \( n \), denoted by \( \langle r_n^k \rangle \), associated to the distribution using the definition provided by the theory of probability

\[ \langle r_n^k \rangle \sim \sum_r r^k T(n|r) \frac{T(n)}{T(n)}. \]  

With this simple simulation we built individual trees and ensembles of trees for animals with sexual reproduction, in order to use them to show characteristics of the process. In the simulation we represented the animal preference for mate by a set of free parameter of the proposed algorithmic model. The examples presented are not intended to be exhaustive, although these examples give an idea about the consequence of parameter selection. This preference could be taken from experimental data, with a different number of individuals of the population under study or different group of original ancestors. Further models could include other characteristics of particular animal groups, even a different way to generate the ensemble of trees.

II.6. About the model computation of sets of trees

Another variable to consider is the time that takes to create a full tree. This variable depends on the length of the tree, longer trees needs more computational time. Simulations up to 100 generations, when we generate \( 2 \times 10^5 \) trees, takes much more than 4 hours. The current code implementation is linear with the generation number. It could be improved in the future using parallel programming tools \[23\]. The model of the tree described here has been implemented in C++ using ROOT libraries \[22\]. This object oriented framework has been used to analyze the results of the simulation too. All tools used in the work correspond to open source packages.

III. REBUILDING THE COMPUTATIONAL MODEL ANALYTICALLY

In this section we describe the simulation analytically, such as a theoretical stochastic process. We will give the tools to the reader in order to describe the algorithm analytically. With them we can describe not only these types of trees, but also we can generalize the process, in order to study this and other problems in this way.
For the kind of trees described in Section II.2 we consider the stochastic process of two random variables \((r_n, s_n)\) on discrete time \(n\). As we said, the random variable \(r_n\) represents the ancestor number at generation \(n\) and the role of the random variable \(s_n\) will be to define paths for the values that can take \(r_{n+1}\), given the acquired value at previous generation, \(r_n\). We essentially distinguish the two random number generators used in the algorithm from the two random variables \((r_n, s_n)\) associated with this analytical description of the algorithm. Although we will continue to use the same notation as the algorithm, by extension.

In Section II.1 we have seen that for \(s_n = 1\) case the sample space of \(r_n, s_n\) is delimited by the obtained value for \(r_{n-1}\), i.e. \(R_n = \{2, 2r_{n-1}\}\). As a general rule, we throw a random number, \(s_n\), at generation \(n\) between two possibilities, for instance \(\{0, 1\}\). According to the result of \(s_n\), we choose one of the following two branches, where \(R_{n+1}^{s_n}\) is the set of values for \(r_{n+1}\) given \(s_n = 0, 1\) and \(r_n\) also. The transition behavior of the process is governed, essentially, by the conditional probability \(P(r_{n+1}|s_n, r_n)\) as we see in Figure 3. In this paper we use different distribution according to the current subprocess.

In Section II.2 we introduce the growth and decay regimes, in this section we describe it analytically as a subprocesses of the main stochastic process. First of all we distinguish this two subprocesses

- **Growth subprocess** \(s_n \in [0, N_\mu]\)

  If \(s_n = 0\), we take \(r_{n+1} = 2r_n\), i.e. we assign the double of the number of ancestors obtained in generation \(n\), as the new number of ancestors for generation \(n+1\). If \(s_n = 1\), we take \(r_{n+1}\) as a random variable, with a certain probability distribution whose support is \([2, 2r_n - 1]\). We use two kind of distribution for \(P(r_{n+1}|r_n, s_n = 1)\) uniform and negative exponential.

- **Decay subprocess** \(s_n \in [N_\mu, N]\)

  The algorithm is similar, but is subdivided according to the generation number \(n\):

  1. **Soft Decay** \(s_n \in [N_\mu, N_\gamma]\). For the branch \(s_n = 0\), we continue with the use of delta distribution that assign for \(r_{n+1}\) the double of the previous generation: \(r_{n+1} = 2r_n\). For the branch \(s_n = 1\), we use only negative exponential distribution for \(P(r_{n+1}|r_n, s_n = 1) \sim e^{-\lambda_n r_{n+1}}\) over the set \(R_{n+1}^1\). The characteristic parameter of this distribution is \(\lambda_n\), it is chosen in order to give more predominance to the smaller values of \(R_{n+1}^1\). Moreover we add a constraint: the value of \(r_{n+1}\) is bounded through the value obtained in \(N_\mu, \rho_n\), this is \(r_{n+1} \leq \rho_n\).

  2. **Hard Decay** \(s_n \in [N_\gamma, N]\). The branch \(s_n = 0\) is changing slightly. Now we choose a random value uniformly distributed between 2 and \(\alpha N\), in order to enforce the decay. This number \(\alpha N\) is pre-fixed and represents the upper bound of the initial number of individuals at the beginning of the population that has made the complete tree. The branch \(s_n = 1\) is governed by the same rule as in the soft decay regime. Also we still use \(r_{n+1} \leq \rho_n\).

  We only use the negative exponential distribution for \(P(r_{n+1}|r_n, s_n)\). Also the value \(r_{n+1}\) is bounded through the value obtained in \(N_\mu, \rho_n\), this is \(r_{n+1} \leq \rho_n\).

    A general diagram in Figure 6 summarizes the sophistication implemented in the algorithm.

**FIG. 6.** Diagram associated to the simulation process. The value of the random variable \(r_{n+1}\) depends of the value obtained in the previous generation for \(r_n\) and \(s_n\). The node \(r_i\) corresponds to the initial condition equal to \(2^{i+1}\) at the initial generation \(i = 0, 1\). The initial condition for \(s_i\) is also a random number on \(\{0, 1\}\), not a prefixed number

For the growth and soft decay subprocesses the random variable \(s_n\) generates a partition of the sample space of \(r_{n+1}\). We see that explicitly from \(R_{n+1}^0 = \{2r_n\}\) and \(R_{n+1}^1 = \{0, 2r_n - 1\}\), then \(R_{n+1} = R_{n+1}^0 \cup R_{n+1}^1\), and \(R_{n+1}^0 \cap R_{n+1}^1 = \Phi\), the empty set.

However, in any regime we will have that

\[
P(r_{n+1}) = \sum_{s_n = 0, 1} P(r_{n+1}, s_n) \tag{6}
\]

because the random variable \(s_n\) defines a particular branch, which involves an exclusive action over \(r_{n+1}\).

In the simulation proposed we choose the same probability distribution for all Bernoulli random variables \(\{s_n\}\); \(P(s_n = 0) = \frac{1}{2}\), for all generations \(n \geq i\). This corresponds to the situation of maximum ignorance (disorder) about the process. We can generalize this situation denoting \(p_n := P(s_n = 1)\) and the complementary probability \(q_n := P(s_n = 0) = 1 - p_n\). Using the definition of conditional probability \(P(r_{n+1}, s_n) = P(r_{n+1}|s_n)P(s_n)\) and from (6) we express \(P(r_{n+1})\) as a convex combination

\[
P(r_{n+1}) = q_n P(r_{n+1}|s_n = 0) + p_n P(r_{n+1}|s_n = 1) \tag{7}
\]

By definition the expected value is

\[
E(r_{n+1}) := \sum_{R_{n+1}} r_{n+1} P(r_{n+1}) \tag{8}
\]
the \textit{special sum} notation defines as follows

\[
\int_{x \in X} \star := \begin{cases} 
\sum_{x \in X} \star : & \text{if } x \text{ numerable}, \\
\int_{x \in X} \star \, dx : & \text{if } x \text{ not numerable}.
\end{cases}
\]

Although we know that the first random number generator \( r_n \), presented in Section 1, takes discrete values. The reason for using a generic notation here, is that it is usually simpler to solve the problem analytically in the continuous case and then to proceed with the discretization of the solution. As we said, we distinguish the random number generator for the simulation of the algorithm from the random variable in the theoretical description of the algorithm.

The expression (7) allows to get an expression for the expected value \( E(r_{n+1}) \) from (8)

\[ E(r_{n+1}) = q_n E(r_{n+1}|s_n = 0) + p_n E(r_{n+1}|s_n = 1), \]

this expression reveals the contributions from each branch to the expected value.

The simulation can be mathematically reformulated as follows: the random number obtained in generation \( n+1 \), \( r_{n+1} \), is affected by the result of two random values in generation \( n \): \( (r_n, s_n) \). We say that in each generation the random variable \( r_n \) run over a certain support \( R_n \). In generation \( n+1 \) the events \( \{(r_n, s_n) : r_n \in R_n, s_n \in \{0, 1\}\} \) conform a partition of the sample space \( R_{n+1} \).

The dependence of the random variable \( r_{n+1} \) with \( r_n \) is implicit in the equations (8) and (9). It appears in the sample space \( R_{n+1} \) explicitly and in the parameters of \( P(r_{n+1}|s_n) \) occasionally. Furthermore the diagram in Figure 3 representing a link in the process, it does not show the probability distribution \( P(r_n) \) explicitly. Nevertheless, in order to provide a more specific description of transition \( r_n \rightarrow r_{n+1} \) we need also to write the probability \( P(r_{n+1}) \) in terms of \( P(r_n) \). The notation is simplified, \( P(r_n) \) means the pdf associated to the random variable \( r_n \).

We use the \textit{Law of total probability} for the general two random variables \((x, y)\)

\[ P(y) = \int_x P(y|x) \, P(x), \]

where \( P(x) \) is a simplification of the notation of the probability to find the random variable \( x \) equal to a certain number, for the discrete case, or in the infinitesimal interval near that includes the value \( x \), for the case of continuous case.

As we show in Figure 7, given this two values \((r_n, s_n)\) there is a certain probability to jump on a particular value \( r_{n+1} \), i.e. if we take \( x \equiv (r_n, s_n) \) and \( y \equiv r_{n+1} \) the total contribution to \( P(r_{n+1}) \) it is ruled by (10)

\[ P(r_{n+1}) = \sum_{s_n=0,1} \int_{r_n \in R_n} P(r_{n+1}|r_n, s_n) \, P(r_n, s_n). \]

In each generation the random variables \( r_n \) and \( s_n \) are independent, then \( P(r_n, s_n) = P(r_n)P(s_n) \), and from (11) we have

\[ P(r_{n+1}) = \int_{r_n \in R_n} T(r_{n+1}|r_n) \, P(r_n), \]

where

\[ T(r_{n+1}|r_n) := \sum_{s_n=0,1} P(r_{n+1}, s_n|r_n) \, 1_{s_n}(r_n), \]

and also \( 1_{s_n}(r_n) \) is an indicator function, defined as \( 1_{s_n}(r_n) = 1 \), if \( r_n \in R_{n,s_n} \) and \( 1_{s_n}(r_n) = 0 \) otherwise. The expression (12) shows the markovian property of this stochastic process. The transition elements \( T(r_{n+1}|r_n) \) and initial condition govern the time evolution of the probability distribution \( P(r_n) \). In Figure 6 the initial arrows connecting the nodes \( s_i \) and \( r_i \) with \( r_{i+1} \) represent the possibility to \textit{throw} a random number \( s_i \), in \( i \)-generation, in order to choose the sample space for the next random number \( r_{i+1} \), given the number \( r_i \). The same sentence is also valid for the arrows connecting the nodes \( s_n \) and \( r_n \) with \( r_{n+1} \). By definition of sexual reproduction specie we have \( r_0 = 2 \). In this work we also assigned \( r_1 = 2r_0 \) directly, without any second random number \( s_0 \) in generation \( n = 0 \). In other words the initial condition for the simulation is given in \( i = 1 \): \( r_1 = 4 \), the \textit{grandparents} generation.

![Figure 7](image-url)
We describe the general algorithm with initial condition at $i$—generation, $i = 0, 1$, denoting by $r_i = 2^{i+1}$, trivially the expected value at this generation is $E(r_i) = 2^{i+1}$.

We can obtain a recurrence equation for growth and soft decay regimes, presented in Section II.3 In that cases the algorithm establishes that

$$P(r_{n+1} | r_n, s_n = 0) = \delta_{r_{n+1}, 2r_n} \quad (14)$$

in the discrete case ($P(r_{n+1} | r_n, s_n = 0) = \delta(r_{n+1} - 2r_n)$, in the continuous case). Using this conditions on all the terms that $s_n = 0$ of expression (11), to reduce the expression (9) and leads

$$E(r_{n+1}) = 2q_nE(r_n) + \xi_{n+1}, \quad (15)$$

this is a non-homogeneous linear recurrence equation, valid from $n \geq i$, where $\xi_{n+1}$ is the in-homogeneity term given by

$$\xi_{n+1} := \int \int_{R_{n+1} \times R_n} r_{n+1}P(r_{n+1} | r_n, s_n = 1) P(r_n) p_n, \quad (16)$$

This equation contains the trivial case: $p_n = 0$ (i.e. $q_n = 1$) for all $n \geq i$, corresponding to no inbreeding in all generations. In this situation the non-endogamic solution of (15) is

$$E(r_n) = 2^{n+1}. \quad (17)$$

We can express the general solution of (15), by iteration until $i$—generation, as follows

$$E(r_n) = 2^{n-i} q_{n-1} \cdots q_i E(r_i) +$$

$$+ \sum_{k=i+1}^{n} 2^{n-k} q_{n-1} \cdots q_{k-1} \xi_k, \quad (18)$$

where $n \geq i + 1$. Again if we take $q_k = 1$ (e.g. $\xi_k = 0$), for all $k$, in (18) we recover the non-endogamic (or trivial) solution (17).

The first term of (18) corresponds to the case in which we have no inbreeding in all previous generations to $n$ (until $i$—generation), because the product $q_i \cdots q_{n-1}$ is just equal to this intersection probability

$$P(s_i = 0, \ldots, s_{n-1} = 0) = P(s_i = 0) \cdots P(s_{n-1} = 0).$$

In the second terms of (18) we have the quotients of $q_k$'s that are just equal to this intersection probability

$$\frac{q_i \cdots q_{n-1}}{q_i \cdots q_{k-1}} = P(s_k = 0, \ldots, s_{n-1} = 0),$$

for $k = 1, \ldots, n - 1$ and is equal to 1, for $k = n$.

The paper simulation is based on the case that $q_n = \frac{1}{2}$, for all $n \geq i$.

$$E(r_n) = E(r_i) + \sum_{k=i+1}^{n} \xi_k, \quad (19)$$

As it was said, that the inhomogeneous terms $\xi_k$ can be written as

$$\xi_k = E(r_k | s_{k-1} = 1)p_n, \quad (20)$$

also we can express (20) as $\xi_k = E(r_k, s_{k-1} = 1)$.

Continuing with the last regime, for the hard decay defined on Section II.3 the algorithm establishes that $r_{n+1}$ is a uniformly distributed in $[2, \alpha_N]$, from (9) we have simply

$$E(r_{n+1}) = q_n \frac{\alpha_N + 2}{2} + \xi_{n+1}. \quad (21)$$

because $E(r_{n+1} | s_n = 0) = (\alpha_N + 2)/2$ at this regime, i.e. is the first raw moment of random variable uniformly distributed in $[2, \alpha_N]$.

On the other hand, we can give more satisfactory description in terms of high order moments of $P(r_n)$. We obtain an expression for growth and soft decay regimes, from the moment generating function associated to $P(r_n)$ we prove that

$$E(r_{n+1}^k) = 2^k q_n E(r_n^k) + r_n^k \xi_{n+1}, \quad (22)$$

where $\xi_{n+1} := E(r_{n+1} | s_n = 1)$. The recurrent equation (22) has the structure as the equation (15) and also share the same kind of solution of equation (18).

For the hard decay regime and (9) we obtain

$$E(r_{n+1}^k) = q_n \sum_{l=0}^{k} 2^k-l \alpha_N + \xi_{n+1}, \quad (23)$$

where $\xi_{n+1} := E(r_{n+1} | s_n = 1)$ and $E(r_{n+1}^k | s_n = 0)$ is the $k$—order raw moment of random variable uniformly distributed in $[2, \alpha_N]$.

We can even translate the center of the moment of $k$—order, assuming that we know all the previous moments $\{E(x^l) : l = 0, 1, \cdots k\}$, using $(x - c)^k = \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} c^{k-l} x^l$, we have

$$E[(x-c)^k] = \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} c^{k-l} E(x^l), \quad (24)$$

This expression can be useful to express $E[(x-E(x))^k]$ as a linear combination of power of $E^n(x)$ and $E(x^n)$ where $n = 0, 1, \cdots, k$, in order to obtain the $k$—order moment centered around the mean value $E(x)$.

This theoretical approach allows to continue with the refinement of free parameters that the algorithm has, observing the analytical behavior of the solution, and in
more general terms, the structure of the evolution equation. Regarding this, we want to add that there are also other ways to conduce the convergence of the tree. Other alternative is to use only a negative exponential distribution \( \sim \lambda_n e^{-\lambda_n n} \) for the whole decay interval \([N_\mu, N]\) and use its parameter \( \lambda_n \) to control the endogamy degree. This alternative exempts us from considering a cut generation, \( N_\gamma \), a priori; there is no need to subdivide the decay interval into soft and hard.

In the following section we perform a comparison between the results of the proposed algorithm and the results of the theoretical paper \[17\].

IV. A COMPARISON WITH THE THEORETICAL TOY MODEL

We used the mean value of the number of ancestors presented in \[17\] to compare it against data obtained with the computational model. The expression for the mean value of the number of ancestors at generation \( n \), namely \( \alpha(n) \), obtain in \[17\] was:

\[
\alpha(n) = 2^n - \beta(n),
\]

where \( \beta(n) \) represents the mean value of individuals who are outside to the set of ancestors, with respect to the maximum possible number of ancestors in each generation \( n \), in this case is equal to \( 2^{n+1} \). Also

\[
\beta(n) = 2^{a+b} \langle X(n) \rangle
\]

is the product of expected value associated to the diluted process, \( \langle X(n) \rangle \), modulated by \( 2^{a+b} \), \[17\]. Explicitly \( \langle X(n) \rangle \) is equal to

\[
\langle X(n) \rangle = e^{-2n}[2n I_1(2n) + (2n + \frac{1}{2})I_0(2n)] - \frac{1}{2n}.
\]

where \( I_0(x) \) are the modified Bessel function \[23\].

The number of ancestors, \( \alpha(n) \), depends on two parameters \( a \) and \( b \). If the expected value satisfies \( \alpha(t_1) = \alpha_1 \) and \( \alpha(t_2) = \alpha_2 \), for two generations \( t_1 \) and \( t_2 \) such that \( t_1 \neq t_2 \), the parameters \( a \) and \( b \) can be obtained by

\[
a = \frac{1}{t_2 - t_1} \log_2 \left[ \frac{2^{t_2+1} - \alpha_2}{2^{t_1+1} - \alpha_1} \frac{\langle X(t_1) \rangle}{\langle X(t_2) \rangle} \right]
\]

\[
b = \frac{1}{t_2 - t_1} \left\{ t_2 \log_2 \left[ \frac{2^{t_2+1} - \alpha_1}{\langle X(t_1) \rangle} \right] - t_1 \log_2 \left[ \frac{2^{t_2+1} - \alpha_2}{\langle X(t_2) \rangle} \right] \right\}
\]

where \( \alpha_i \leq 2^{t_i+1} \), for \( i = 1, 2 \), to ensure good definition of \( a \) and \( b \). These parameters can be related with the maximum number of ancestors in a given generation and the horizontal range \[17\].

We used the equation for the mean number of ancestors depending on these parameters and performed a fit of the expression to our simulation data, in Figure 8.

Also it is possible to obtain \( a \) and \( b \) parameters for any case that we want to study. The values for the fit parameter will make sense when experimental data would be use.

The Table I summarizes the values for \( a \) and \( b \) obtained for different distributions and the same final generations to end the process.

The mean value of the number of ancestors in the set behaves with the generation in a similar way to the mean value obtained the theoretical model. This simulation could include more specific information of particular species or field studies with animals. Also it could be possible to combine the simulation with genetic algorithms to obtain a powerful tool to trace the combination of genes through the history of a particular specie.
TABLE I. Values of the fit parameters for two different cases corresponding to Figure 5 for trees of maximum generation $N = 20$.

| Distribution               | $a$    | $b$    | $\pm 0.01$ | $\pm 0.004$ |
|----------------------------|--------|--------|-------------|-------------|
| Uniform distribution       | 0.997  | -0.73  |             |             |
| Negative Exponential       | 0.956  | -0.25  |             |             |
| distribution               |        |        |             |             |

Even when we can obtain close results fitting the theoretical model parameters, our simulation is an improvement on the first theoretical model, in the sense that it is possible to modify and chose a certain distribution to model animal preferences in mate selection. Also the case studied in [17], that is a theoretical model in particular conditions, could not reproduce a slow decreasing of ancestors number as it is expected in a soft transition in consecutive generations. As we said in [17] the result of this model can be improved with a different choice of the gauge function $\lambda(t)$. In this algorithmic model, we can simulate an specific behavior in animal mate preference, according to the branch probability $q_n$ and the transition probabilities $P(r_{n+1}|s_n, s_n)$. This algorithm propose a more robust model that the first one proposed in [17].

V. CONCLUSIONS AND FURTHER WORK

In present work we have built a simulation that is ruled by a small number of parameters and generates trees of ancestors. This simulation is based on a recursive algorithm and a random number generator. The model presented allows us to include animal mate preferences and to build more realistic trees than the previous model presented in [17].

This more sophisticated model allows us to include biological considerations represented through the distribution and its parameters.

Again meaning different reproductive behavior such as one male mating with several females in a group or sibling selection, could be changed via the percent rate in this second generator $s_n$ from Section [11.2]

Even when this model has many empirical elements, it is a more realistic simulation of the trees that our previous version. Also in this regard, the simulation presents an opportunity to explore and discuss the elements of the mathematical description in a process of stochastic nature, in this case with markovian properties.

To develop the simulation, we used the available tools learned from the field of physics to generate a flexible dynamical model that could be used for biologist to compare and make predictions with real animal data.

The model mainly uses the hypothesis of a certain degree of inbreeding as the key to the development of any animal population of sexual reproduction. We leave open the question of which is the degree of endogamy required to a healthy population development but we claim that no population could develop without an certain level of endogamy within.

Our model could be used as a powerful tool in order to contribute in ecology and biology studies by using empirical data collected from an animal behavior of a any population to constraint the parameters of the model and make predictions. Besides the algorithm applied on this simulation can be used to describe other biological or physical systems with similar dynamics. Statistical models of biparental reproduction have already been compared with physical systems, such as spin-glass systems [20]. Also in this regard the evolutionary graph theory is an approach to study how topology affects the evolution of a population [27].

Some open questions related to the nature of the process have raised from the developed simulation, for instance: How many generations makes sense to follow a tree of ancestors? How to chose where to stop the tree? What happens with the human case? Could be possible with this model shows differences between animal groups of sibyl species.

We have developed a model through an algorithm that allows to gain understanding of the future experiment. We show how relatively simple selections of this two distributions: the one-dimensional $P(s_n)$ and the conditional $P(r_{n+1}|s_n, r_n)$ allow us to describe the beginnings of a phenomenology of the concrete process involved in the ancestors trees formation. Such phenomenology can be enriched as long as these probability distributions become more complex.

Further work will include a development of an algorithm version in python code, with parallel programing improvements.

ACKNOWLEDGMENTS

This work was supported by the following institutions: CONICET and UNQ. We thank the scientists from the biology field who encourage us to keep working in biological models. We thank also our reviewers Graciela Molina and Pablo Alcain for their contribution to this work. Special mention to Cristina J. for his critical objections and text style corrections. Finally, thanks mention to Micaela Moreton, María Clara Caruso, and Gabriel Lio for always give us personal support. During the development
of this work we have a member to include in the Caruso family tree (little Lucia) and we want to dedicate present work to her.

[1] Ohno, S. The Malthusian parameter of ascents: What prevents the exponential increase of one's ancestors? Proc. Natl. Acad. Sci. 93, 1527615278 (1993)
[2] Charlesworth, D., and Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. Annu. Rev. Ecol. Sys. 18, 237268.
[3] Peer, K., and Taborsky, M. (2005). Outbreeding depression, but no inbreeding depression in haplodiploid ambrosia beetles with regular sibling mating. Evolution Int. J. Org. Evolution 59, 317323.
[4] Templeton, A.R. (1986). Coadaptation and outbreeding depression. In Conservation Biology: The Science of Scarcity and Diversity, M.E. Soule, ed. (Sunderland, MA: Sinauer Associates), pp. 105116
[5] Pusey, A., and Wolf, M. (1996). Inbreeding avoidance in animals. Trends Ecol. Evol. 11, 201206.
[6] Gerlach, G., and Lysiaik, N. (2006). Kin recognition and inbreeding avoidance in zebrafish Anim. Behav. 71, 13711377.
[7] Frommen, J.G., and Bakker, T.C.M. (2006). Inbreeding avoidance through non-random mating in sticklebacks. Biol. Lett. 2,232235.
[8] Bateson, P. (1983). Optimal outbreeding. In Mate Choice, P. Bateson, ed. (Cambridge: Cambridge University Press), pp.257277.
[9] Bateson, P. (1982). Preferences for cousins in Japanese quail., Nature 295, 236237.
[10] Neff, B.D. (2004). Stabilizing selection on genomic divergence in a wild fish population., Proc. Natl. Acad. Sci. USA 101, 2381-2385.
[11] Willi, Y., and Van Buskirk, J. (2005). Genomic compatibility occurs over a wide range of parental genetic similarity in an outcrossing plant. Proc. R. Soc. Lond. B. Biol. Sci. 272, 13331338.
[12] Jamieson, I.G. et al. (2009) Why some species of birds do not avoid inbreeding: insights from New Zealand robins and saddlebacks. Behav. Ecol. 20, 575584.
[13] Rioux-Paquette, E. et al. (2010). No inbreeding avoidance in an isolated population of bighorn sheep. Anim. Behav. 80, 865871.
[14] Szulkin, M. et al. (2009) Inbreeding avoidance under different null models of random mating in the great tit. J. Anim. Ecol. 78, 778788.
[15] Thunken, T. et al. (2007) Active inbreeding in a cichlid fish and its adaptive significance. Curr. Biol. 17, 225229.
[16] Nichols HJ, Cant MA, Hoffman JL, Sanderson JL. 2014. Evidence for frequent incest in a cooperatively breeding mammal. Biol. Lett. 10: 20140898.
[17] M.Caruso, C. Jarne. Physical Review E 90, 022125 (2014).
[18] J. H. Bennett, Ann. Eugen. 17, 311 (1952).
[19] J. H. Bennett, J. Theor. Biol. 4, 28 (1963).
[20] G. H. Hardy, Sci. N. S. 28, 49 (1908).
[21] W. Weinberg, U ber den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl Naturkd. Wurttemb. 64, 369 (1908) (English translations in Boyer 1963 and Jamenson 1977).
[22] Rene Brun and Fons RadeSmakers, ROOT - An Object Oriented Data Analysis Framework, Proceedings AIHENP’96 Workshop, Lausanne, Sep. 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. See also http://root.cern.ch/
[23] C. Jarne. Markov process simulation for ancestors trees (the code parallelization). Talk from the Introductory School on Parallel Programming and Parallel Architecture for High-Performance Computing, Trieste; 2016. http://indico.ictp.it/event/7659/session/19/contribution/95