Accepted Manuscript

Registering the evolutionary history in individual-based models of speciation

Carolina L.N. Costa, Flavia M.D. Marquitti, S. Ivan Perez, David M. Schneider, Marlon F. Ramos, Marcus A.M. de Aguiar

PII: S0378-4371(18)30711-8
DOI: https://doi.org/10.1016/j.physa.2018.05.150
Reference: PHYSA 19690

To appear in: *Physica A*

Received date : 10 June 2017
Revised date : 10 April 2018

Please cite this article as: C.L.N. Costa, F.M.D. Marquitti, S.I. Perez, D.M. Schneider, M.F. Ramos, M.A.M. de Aguiar, Registering the evolutionary history in individual-based models of speciation, *Physica A* (2018), https://doi.org/10.1016/j.physa.2018.05.150

*Highlights (for review)

Reply Ms. Ref. No.: PHYSA-171117

Title: Constructing phylogenetic trees in individual based models

Physica A

Highlights

- We provide a link between individual-based models and macroevolutionary theory.
- We show how to track ancestral relationships and speciation/extinction events in IBMs.
- Genealogies of individuals and phylogeny of species are drawn from these algorithms.
- We illustrate these algorithms using a spatially-explicit model of speciation.
- We compare trees based on historical information with trees inferred from genetic data.

Registering the evolutionary history in individual-based models of speciation

Carolina L. N. Costa[a,*], Flavia M. D. Marquitti[b], S. Ivan Perez[c,b], David M. Schneider[b], Marlon F. Ramos[b], Marcus A.M. de Aguiar[a,b]

[a] *Instituto de Biologia, Universidade Estadual de Campinas, Unicamp, 13083-859, Campinas, SP, Brazil*
[b] *Instituto de Física 'Gleb Wataghin', Universidade Estadual de Campinas, Unicamp, 13083-859, Campinas, SP, Brazil*
[c] *División Antropología, Museo de La Plata, Universidad Nacional de La Plata, Paseo del Bosque s/n, 1900 La Plata, Argentina*

Abstract

Understanding the emergence of biodiversity patterns in nature is a central problem in biology. Theoretical models of speciation have addressed this question in the macroecological scale, but little has been done to connect microevolutionary processes with macroevolutionary patterns. Knowledge of the evolutionary history allows the study of patterns underlying the processes being modeled, revealing their signatures and the role of speciation and extinction in shaping macroevolutionary patterns. In this paper we introduce two algorithms to record the evolutionary history of populations and species in individual-based models of speciation, from which genealogies and phylogenies can be constructed. The first algorithm relies on saving ancestor-descendant relationships, generating a matrix that contains the times to the most recent common ancestor between all pairs of individuals at every generation (the Most Recent Common Ancestor Time matrix, MRCAT). The second algorithm directly records all speciation and extinction events throughout the evolutionary process, generating a matrix with the true phylogeny of species (the Sequential Speciation and Extinction Events, SSEE). We illustrate the use of these algorithms in a spatially explicit individual-based model of speciation. We compare the trees generated via MRCAT and SSEE algorithms with trees inferred by methods that use only genetic distance between individuals of extant species, commonly used in empirical studies and applied here to simulated genetic data. Comparisons between trees are performed with metrics describing the overall topology, branch length distribution and imbalance degree. We observe that both MRCAT and distance-based trees differ from the true phylogeny, with the first being closer to the true tree than the second.

Keywords: genealogies of individuals, phylogenies of species, macroevolutionary patterns, distance-based trees, tree statistics

*Corresponding author. Phone: +55-19-35215466
Email address: lemes.carol@gmail.com (Carolina L. N. Costa)

1. Introduction

The origin of the patterns of diversity at macroecological scale is a central problem in biology [1–3]. In the last decades patterns such as geographical variation in species richness, species abundance distributions and species-area relationships, have been studied from empirical and theoretical perspectives [4–8]. Neutral models of speciation – where differences between individuals are irrelevant for their birth, death, and dispersal rates [3, 9] – have played a central role in understanding the patterns of diversity at the macroecological scale. With the help of computers, it became possible to test different hypothesis about the mechanisms of speciation, such as sympatric versus allopatric processes, assortative mating and the effect of number of genes [10–12].

Among the different theoretical approaches designed to quantitatively study speciation [3, 13], models that explicitly incorporate space have allowed the study of major macroecological patterns that could be compared with those observed in nature [2, 7, 14, 15]. However, these models have given little attention to the historical or evolutionary dimension of the origin of diversity, which is reflected in the macroevolutionary patterns described by phylogenetic trees [16–19]. Because of the increased interest in the role of microevolutionary processes on the resulting macroecological patterns, the extension of these approaches to include algorithms that track the branching or phylogenetic divergence process is a next fundamental step to further explore models of speciation using simulations [16, 20, 21]. Individual-based models (IBM) widely used in biology [22] have the advantage that can be easily extended to include this historical perspective and to provide a record of the ancestor-descendant relationships among the simulated individuals and/or species. These relationships can be stored in matrices from which individual genealogies and species trees (i.e. phylogenies) may be directly obtained.

In this article we describe two algorithms that save historical information in individual-based models of speciation. The first algorithm focuses on genealogies and the quantity saved is the parenthood of each individual. With parenthood registered, the *time to the most recent common ancestor*, i.e., the number of generations needed to go backward to find a common ancestor of one individual with another individual of the population, can be easily calculated in terms of the common ancestor of the parents. These times are computed at every generation between all pairs of individuals and, at the end of the simulation, are saved in a matrix (the Most Recent Common Ancestor Time matrix - MRCAT). The second algorithm focuses on phylogenies and consists of directly records all speciation and extinction events (the Sequential Speciation and Extinction Events - SSEE) and set a matrix analogous to MRCAT but whose entries are species rather than individuals. The SSEE matrix contains the exact branching times in the simulated clade or community, including all extinct species. The MRCAT and SSEE matrices can be used to draw the exact branching sequence of the simulated individuals and species, respectively. These procedures differ from the inference methods based on phenotypic and genetic traits used to estimate phylogenies in natural studies, because in our model we are looking

for the branching process forward in time, while in usual approaches the same process is looked backwards in time. In addition to the presentation of the MRCAT and SSEE algorithms, we compare the trees they generate with those obtained by usual distance-based methods of phylogenetic inference using only genetic data from simulated individuals of the final community. Comparing these inferred phylogenies with those generated by MRCAT or SSEE algorithms might offer a practical way to evaluate the reliability of the estimated trees to recover natural macroevolutionary patterns.

The paper is organized as follows: in section 2 we describe the algorithms to record ancestor-descendant relationships (MRCAT, subsection 2.1) and speciation/extinction events (SSEE, subsection 2.2). In subsection 2.3 we compare the true phylogenetic tree obtained from the SSEE algorithm with genealogies of individuals obtained from the MRCAT algorithm considering only one individual per species. In section 3 we discuss the applications of the algorithms proposed in section 2. First, we present an individual-based model of speciation proposed in [2] in which the algorithms regarding the ancestor-descendant relationships and the branching process were incorporated (subsection 3.1). We emphasize that the algorithms are quite general and could be implemented in most IBM's. Next, we briefly describe the Unweighted Paired Group Method with Arithmetic mean (UPGMA) [23], the Neighbor Joining (NJ) [24] and the Minimum Evolution (ME) [25] methods, which are based on genetic distances calculated directly from one individual of each species present in the last generation of the simulation (subsection 3.2). While closer to what empiricists do, the phylogenies derived from these methods are further from the true phylogeny generated by the SSEE algorithm than is the phylogeny based on the MRCAT algorithm presented here. We end this section presenting the statistical measurements used to compare phylogenies obtained from algorithms proposed here with those estimated by distance-based methods (subsection 3.3). The goal is to show that the accuracy of some methods usually employed when the only information available is the data of individuals collected from nature can be evaluated with the help of models. In section 4 we present the results regarding the output of simulations and the comparisons of phylogeny summary statistics. Finally, section 5 was devoted to discussion and section 6 to conclusions.

2. Registering the history of individuals and species

In this section we describe two algorithms to record historical information during the evolution of a population. The first algorithm records genealogical relationships between all pairs of individuals at every generation. The second, in turn, registers all the speciation and extinction events that occur along the evolutionary history. These algorithms are general enough to be applied to most individual-based models of speciation.

2.1. Ancestor-descendant relationships among individuals - MRCAT

In this subsection we show how the time to the most recent common ancestor between all pairs of individuals can be obtained by keeping track of parental re-

Individuals at generation $t+1$	Parent at generation t
1	$P(1) = 4$
2	$P(2) = 8$
3	$P(3) = 1$
4	$P(4) = 4$
...	...
N_{t+1}	$P(N_{t+1}) = 15$

Table 1: List of individuals (i) at generation $t+1$ and their respective parents ($P(i)$) at generation t in an asexual model. This information is necessary to construct the MRCAT matrix. Parents of each individual must be recorded to track the most recent common ancestor between individuals at the end of a simulation. Note that individuals at generation t are not the same individuals at generation $t+1$ (discrete generations).

lationships at every generation. We also show how this information can be used to draw the genealogy of individuals of the last simulated generation. We distinguish between asexual and sexual models because of the technical differences in tracking only one or two parents.

2.1.1. Asexual models

Consider a population of N_t asexual individuals at generation t. The population at the next generation, $t+1$, will be comprised of offspring of these individuals and the parent of individual i will be denoted $P(i)$.

An example is shown in Table 1, where $P(1) = 4$, $P(2) = 8$, $P(3) = 1$, etc. The MRCAT between individuals i and j is

$$T_{t+1}(i,j) = T_t(P(i), P(j)) + 1. \tag{1}$$

which is simply the time to the most recent common ancestor between the parents plus one, since a generation has passed [26]. As examples

$$T_{t+1}(1,2) = T_t(4,8) + 1$$

and

$$T_{t+1}(1,4) = T_t(4,4) + 1 = 1.$$

since in this last case they have the same parent. Starting from $T_0(i,j) = 1$ if $i \neq j$ and noting that $T_t(i,i) = 0$ at all times the rule (1) allows one to compute the MRCAT matrix for any number of generations. The matrix T is stored only for two times, the past and the present generation, so that the memory cost does not depend on time, only on the (square) size of population. A schematic view of the algorithm is shown in Fig. 1, where the genealogical relationships between 9 individuals originated from a single ancestor is represented. In this example the total population size is kept fixed, so that the full MRCAT matrix is always 9×9. The phylogeny of the community can be drawn by selecting one individual per species at each moment in time. The corresponding matrices at $t = 3$ and $t = 6$ are given by

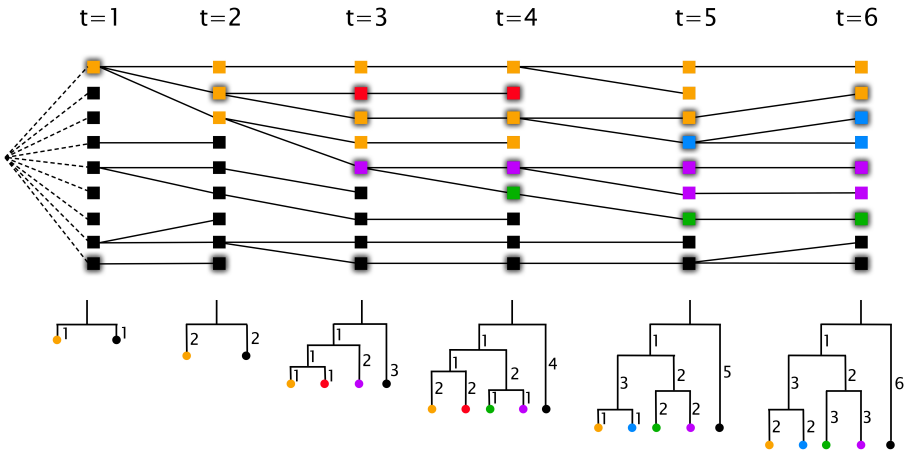

Figure 1: Illustration of ancestor-descendant relationships for an asexual population with constant size $N = 9$ implemented with MRCAT algorithm. Each square is an individual and colors represent different species. Phylogenetic trees are constructed by selecting one individual per species (shaded squares).

$$T_3 = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 2 & 3 \\ 2 & 2 & 0 & 3 \\ 3 & 3 & 3 & 0 \end{pmatrix}; \qquad T_6 = \begin{pmatrix} 0 & 2 & 5 & 5 & 6 \\ 2 & 0 & 5 & 5 & 6 \\ 5 & 5 & 0 & 3 & 6 \\ 5 & 5 & 3 & 0 & 6 \\ 6 & 6 & 6 & 6 & 0 \end{pmatrix}. \quad (2)$$

where the selected individuals are shown in shaded colors (from top to bottom) at the corresponding times.

2.1.2. Sexual models

The generation of MRCAT matrices in sexual models is slightly different, since each individual i has two parents, a mother $P_1(i)$ and a father $P_2(i)$. Consider as an example a population which has 4 females and 3 males in generation t and gives rise to 5 females and 3 males in generation $t + 1$ (Table 2). Notice that not only the total number of individuals but also the number of males and females may vary over generations. As the model is sexual, both maternal and paternal lineages can be followed in the simulations, allowing the generation of two different MRCAT matrices and their corresponding trees. A third option is not tracking lineages by sex, but record the most recent common ancestor taking into account both parents, which is the only option if the model considers hermaphroditic individuals.

– *Maternal and paternal lineages.* The maternal lineage of individuals is obtained by computing the time to the most recent common ancestor of their

Individuals at generation $t+1$	Mother at generation t	Father at generation t
Females		
1	$P_1(1) = 4$	$P_2(1) = 6$
2	$P_1(2) = 3$	$P_2(2) = 7$
3	$P_1(3) = 1$	$P_2(3) = 7$
4	$P_1(4) = 4$	$P_2(4) = 5$
5	$P_1(5) = 2$	$P_2(5) = 6$
Males		
6	$P_1(6) = 1$	$P_2(6) = 5$
7	$P_1(7) = 3$	$P_2(7) = 5$
8	$P_1(8) = 3$	$P_2(8) = 7$

Table 2: List of individuals (i) at generation $t+1$ and their respective parents ($P_1(i) = mother$ and $P_2(i) = father$) at generation t in a sexual model. In this case each individual has two parents, P_1 and P_2. Notice that the couple 3 and 7 at generation t had two offspring, the individuals 2 and 8 at generation $t+1$, while other couples had only one offspring. Additionally, notice that there were 4 females and 3 males at generation t, while there are 5 females and 3 males at generation $t+1$.

corresponding mothers:

$$T^M_{t+1}(i,j) = T^M_t(P_1(i), P_1(j)) + 1 \tag{3}$$

with $T^M_0(i,j) = 1$ if $i \neq j$ and $T^M_t(i,i) = 0$. Similarly, the paternal lineage is computed with

$$T^F_{t+1}(i,j) = T^F_t(P_2(i), P_2(j)) + 1 \tag{4}$$

with $T^F_0(i,j) = 1$ if $i \neq j$ and $T^F_t(i,i) = 0$. Both T^M and T^F are computed for all individuals, females and males.

– *Lineages of hermaphroditic individuals.* Many simulations consider, for simplicity, hermaphroditic individuals. In this case, the separation into maternal and paternal lineages does not make sense and the definition of the MRCAT matrix is

$$T_{t+1}(i,j) = \min_{\{k,l\}}\{T_t(P_k(i), P_l(j))\} + 1 \tag{5}$$

with $k, l = \{1, 2\}$, $T_0(i,j) = 1$ and $T_t(i,i) = 0$. This considers, literally, the most recent common ancestor of i and j, taking all parental combinations into account. The same definition is applied to sexual models with sex separation when the recorded genealogy does not separate the maternal and paternal lineages. In the case of hermaphroditic model the MRCAT matrix does not determine the tree uniquely. A detailed example of this situation is described in Supporting Information, section I.

2.1.3. Drawing genealogies from MRCAT matrices

At the end of the simulated evolutionary process the MRCAT matrix contains the time to the most recent common ancestor between every pair of individuals of the extant population and this information can be used to draw

genealogical trees. Drawing the tree from the MRCAT matrix consists in joining individuals into groups according to their most recent common ancestral (Fig. 1). The tree starts with N *units* (the extant individuals) and at each step of the process two of these units are joined together to form a group, so that the number of units decreases by 1. Next, the time to the most recent common ancestral between the newly formed group and the other units of the tree (previously formed groups or extant individuals) are recalculated with a so called *clustering method*. Once the times have been recalculated, the pair of units with the least time is joined into a new group. The process ends when a single unit is left, the root of the tree. As discussed in the SI, section I, a unique tree is generated independently of the clustering method for asexual, maternal or paternal lineages. For hermaphroditic populations or for sex separation but with the MRCA taking into account both parents that is not the case. In these situations more than one tree can be constructed from the same MRCAT matrix using different clustering procedures. In all cases the tips (or leaves) of the tree represent extant individuals whereas internal nodes represent the most recent common ancestor between a pair of individuals. Branch length denote the time in generations between an ancestor and its descendants (see, for instance, Fig. S1 in the SI). More information about the drawing of trees is available in Supporting Information, section II.

2.2. Recording all speciation and extinction events - SSEE

The algorithm described in subsection 2.1 records the ancestor-descendant relationships between all pairs of individuals in the population at a given point in time. This allows the drawing of entire genealogies. However, information about individuals that died without leaving descendants or species that went extinct is totally lost. In this subsection we describe an algorithm that allows the construction of the true phylogenetic tree, retaining information about all species that ever existed during the evolution (Fig. 2).

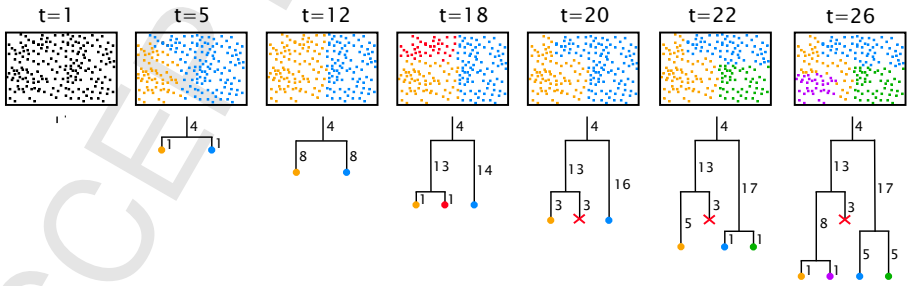

Figure 2: Illustration of speciation and extinction events implemented with SSEE algorithm and the corresponding phylogenetic trees exhibiting the complete history. Colored squares represent individuals of different species, and colored circles in phylogenies represent each species, with numbers denoting the time to speciation and extinction events.

We will use a new matrix S_t (the SSEE matrix) such that $S_t(i, j)$ is the time when species i and j branched off a common ancestral species. Species that

go extinct will be kept in the matrix but will be assigned a label to distinguish them from living (extant) species. This label will be stored in a *extinction vector* E_t such that $E_t(i) = 0$ indicates a living species at time t and $E_t(i) = \tau \neq 0$ indicates the moment τ when the species disappeared.

The algorithm is as follows: consider the hypothetical sequence of speciation and extinction events displayed in Fig. 2. At time t=18 there are three species that we denote as Orange(18), Red(18) and Blue(18) and the corresponding S matrix and E vector are

$$S_{18} = \begin{pmatrix} 0 & 1 & 14 \\ 1 & 0 & 14 \\ 14 & 14 & 0 \end{pmatrix}; \qquad E_{18} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \tag{6}$$

Two generations later, at $t = 20$, one finds only two species, Orange(20) and Blue(20). Notice that names (and colours) are arbitrary and to determine the relation between these species and the ones at the previous time step we need to look at the parents of individuals in each species. Suppose, as illustrated in the figure, that we find that the parents of individuals in Orange(20) belonged to species Orange(18). In this case we draw a link between Orange(18) and Orange(20) and mark Orange(18) as a species that survived that time step, i.e., we set $E_{20}(1) = 0$. Similarly Blue(20) links with Blue(18) and $E_{20}(2) = 0$. Looking at the previous generation we notice that species Red(18) did not leave any descendant species, i.e., it went extinct. In order to keep track of it we create a virtual species Red(20) and set $E_{20}(3) = 20$ as a mark that it is no longer a living species and went extinct at time 20. The SSEE and E vector at time 20 become

$$S_{20} = \begin{pmatrix} 0 & 16 & 3 \\ 16 & 0 & 16 \\ 3 & 16 & 0 \end{pmatrix}; \qquad E_{20} = \begin{pmatrix} 0 \\ 0 \\ 20 \end{pmatrix}. \tag{7}$$

Extinct species are, therefore, treated as species that will never again speciate, but will be kept in the matrix. When drawing the corresponding tree its branch will stop at the value $E(i)$. Proceeding in this way, with the living species always filling the first part of the matrix, followed by copies of extinct species, we can draw the complete phylogeny and study extinction dynamics as well. At time $t = 26$ the SSEE matrix and extinction vector E are

$$S_{26} = \begin{pmatrix} 0 & 1 & 22 & 22 & 9 \\ 1 & 0 & 22 & 22 & 9 \\ 22 & 22 & 0 & 5 & 22 \\ 22 & 22 & 5 & 0 & 22 \\ 9 & 9 & 22 & 22 & 0 \end{pmatrix}; \qquad E_{26} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 20 \end{pmatrix}. \tag{8}$$

One important case occurs when two species merge into a single species (speciation reversal). This might happen, for instance, when two species that have just become reproductively isolated are able to breed again because of a mutation. The resulting merged species will have individuals with parents in

both ancestral species and we need to define which one "survived" and which went extinct. Although this is just a matter of labeling the species, we call the surviving species the one with most parents in the previous generation.

The drawing of species phylogenies for SSEE matrices is almost identical to that for MRCAT matrices. The only differences are that internal nodes represent speciation events, not the time to MRCA, and branches associated to extinct species should not be drawn all the way down to present time, but should stop at the extinction time recorded in the vector E. As in the MRCAT case of separation of lineages by sex, a unique tree is generated independently of the clustering procedure chosen, due to the exact times of speciation and extinction recorded in simulations based on this algorithm.

2.3. Phylogenies generated by ancestor-descendant relationships (MRCAT) versus trees from speciation and extinction events (SSEE)

At the end of a simulation the MRCAT matrix contains the exact time to the most recent common ancestor between every pair of individuals in the population. The SSEE matrix contains the equivalent information at the species level, including extinct species. Both matrices can be used to draw phylogenetic trees. To draw a phylogeny of species considering the ancestor-descendant relationships between individuals we can use the MRCAT matrix with the following reasoning: if N_S species exist at time t and $ind(i,j)$ is the j-th individual of the i-th species, a $N_S \times N_S$ sub-matrix of the full MRCAT matrix can be generated considering only one individual per species (Fig. 1); a simple choice is to take $ind(i,1)$ for $i = 1, 2, \ldots N_S$ so that $T^{phy}_{i,j} \equiv T_{ind(1,i),ind(1,j)}$.

The tree drawn from the SSEE algorithm is the true phylogeny of species, because it records the exact speciation and extinction events, representing the actual branching process. On the other hand, the phylogeny of species drawn from the MRCAT algorithm is different, although similar, from the true phylogeny, because the time to the most recent common ancestor between individuals of different species is only an approximation to the speciation time, since speciation can happen several generations later. Figure 3 illustrates this situation: if a population splits into three species in two closely spaced speciation events, it might happen that the first group to speciate, species A in the figure, has a more recent common ancestor with the subgroup B than B with C. During the time when B and C still form a single species reproduction between their individuals might not happen for a while until they split, preserving the long time ancestry. This is more likely to happen in populations with a spatial structure when individuals belonging to the two subpopulations occupy different areas.

3. Applications of MRCAT and SSEE algorithms to an individual-based model

3.1. The speciation model

The model considered here to exemplify the MRCAT and SSEE algorithms is an extension of the speciation model introduced in [2] and adapted in [27] to

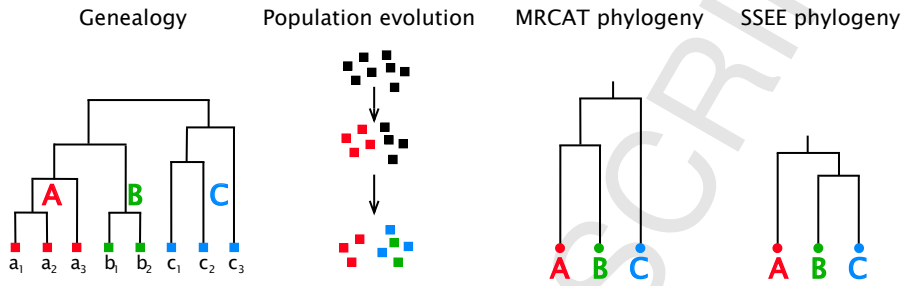

Figure 3: Illustration of a genealogy recorded with MRCAT and the corresponding population evolution. The phylogenies constructed via MRCAT and SSEE differ in this case because, although individuals from species A and B have a more recent common ancestor than with individuals in C, species A split first, followed by the separation of B and C.

characterize individuals with separated sexes (males and females). The model has already been studied in terms of speciation rates, species-area relationships and species abundance distributions. Here we are adding the historical information generated by MRCAT and SSEE algorithms, i.e., recording the parenthood of individuals from one generation to another (genealogy) as well as the pattern and time of the speciation and extinction events (phylogeny or time tree).

The model describes a population of N haploid individuals that are genetically identical at the beginning of the simulation and are randomly distributed in a $L \times L$ spatial lattice with periodic boundary conditions. More than one individual is allowed in each site of the lattice, but because the density of the population is low, this seldom occurs. The genome of each individual is represented by a sequence of B binary *loci*, with state 0 or 1, where each *locus* plays the role of an independent biallelic gene. Individuals also carry one separate label that specify their sex, male or female. The evolution of the population involves the combined influence of sexual reproduction, mutation and dispersal [2].

The reproduction trial starts with individual 1 and goes to individual N, so that all individuals of the population have a chance to reproduce. The individual selected for reproduction, the *focal individual*, searches for potential mates in its *mating range*, a circular area of radius S centered on its spatial location. The focal individual can only reproduce with those within its mating range and if they are genetically compatible, i.e., if the genetic distance between them is below a particular threshold G. Among the compatible individuals within its mating range one of the opposite sex is randomly chosen as mating partner. Individuals whose genetic distance is larger than G are considered reproductively isolated (threshold effect [3]). Genetic distances between individuals are calculated as the Hamming distance [28] between their genetic sequences, i.e., the number of *loci* at which the corresponding alleles are different.

Once the focal individual finds a compatible mate of the opposite sex, reproduction proceeds with the combination of their genetic materials to produce the offspring genome, with each *locus* having an equal probability of being transmit-

ted from mother or father. After combination of parental genomes, each *locus* in the offspring genome can mutate with probability μ. Finally, the offspring replaces the focal reproducing individual. In each reproductive event only one descendant is generated. The offspring is then dispersed with probability D to one of the 20 nearest sites (radius approximately equal to $\sqrt{5} \approx 2.24$) around the expiring focal parent. Conversely, with probability $1 - D$ the offspring will be placed exactly in the same site of its focal expiring parent. Hence, close to the location of every individual of the previous generation there will be an individual in the present generation, keeping the spatial distribution homogeneous. There is a probability Q that the focal individual will die without reproducing. In this case a neighbor is randomly selected from its mating range to reproduce in its place, so that the population size remains constant.

Evolution proceeds in non-overlapping discrete generations such that the entire population is replaced by offspring. Species are defined as groups of individuals connected by gene flow, so that any pair of individuals belonging to different species are reproductively isolated (genetic distance greater than G). However, two individuals belonging to the same species can also be reproductively isolated, as long as they can exchange genes indirectly through other individuals of the species. This model is considered neutral because individuals choose their mates randomly from a mating range, independent of their genetic composition except for the genetic threshold of reproductive compatibility, so differences between individuals are irrelevant for their birth, death, and dispersal rates [3, 9].

3.2. Phylogenies based on genetic distances

As we have described in the previous subsection, the genome of all individuals are identical at the beginning of the simulation but mutations introduce differences and after many generations the population will display a distribution of genomes. Genetic distances can, therefore, be calculated between pairs of individuals and be used as a proxy for ancestry, such that the larger the genetic distance between two individuals the farther back should be their common ancestor. In order to estimate phylogenies by genetic distance, we selected the same individuals per species that were used to draw the phylogeny via MRCAT and computed a matrix of genetic distances. This process mimics the sampling of individuals from a real population and the comparison of their DNA's as a measure of ancestry.

From the genetic distance matrix, we estimated trees from three distance-based methods. Firstly, we used the UPGMA hierarchical clustering method [23]. In this algorithm two groups of species are clustered based on the average distance between all members of the groups. This method assumes a constant rate of change, generating ultrametric trees in which distances from the root to all tips are equal. Secondly, we used the NJ method [24] of phylogenetic inference. In this method the procedure is to find pairs of neighbors in which the total branch length at each stage of the clustering is minimal, starting with a starlike tree. Finally, we used the ME method [25], which assumes that the true phylogeny is probably the one with the smallest sum of branch lengths, as in the

NJ method. The difference is that in the ME method a NJ tree is constructed first and next tree topologies close to this NJ tree are estimated by certain criteria, with all these trees being examined and the tree with the small sum of branch lengths being chosen. We used the function `hclust` of the `stats` package in R [29] to estimate ultrametric trees from the UPGMA method. To estimate trees from the NJ method, we used the `nj` function of the `ape` package in R [30]. In this case, the estimated trees are not ultrametric, so we transform then in ultrametric trees using the `chronoMPL` and `multi2di` functions in `ape` package [30, 31]. We used the `Rkitsch` function of the `Rphylip` package in R [32, 33] to estimate ultrametric trees from the ME method assuming an evolutionary clock. The NJ and ME methods are generally considered superior to UPGMA because they optimize a tree according to minimum evolution criteria. Similarly to the UPGMA, the NJ and ME methods are fast and efficient computationally.

3.3. Statistical indexes to compare phylogenies

To evaluate the accuracy of the phylogenies generated by the MRCAT algorithm and by the genetic distance methods (UPGMA, FM and ME) in relation to the true phylogeny generated by SSEE we use three statistics: the Robinson and Foulds (RF [34]) metric, the gamma statistic (γ [35]) and the Sackin's index (I_s [36, 37]).

The RF metric measures the distance between phylogenetic trees, providing the overall topological resemblance of the phylogenies. Specifically, the RF metric calculates the number of internal branches present in only one of the trees being compared. Given two trees, $T1$ and $T2$, we define

$$RF(T_1, T_2) = \frac{L_1}{L_1'} + \frac{L_2}{L_2'} \tag{9}$$

where L_1 and L_2 are the number of branches on T_1 and T_2, respectively. The number of branches shared by T_1 and T_2 are represented by L_1' and L_2'. The RF metric was calculated using the `RF.dis` function of the `phangorn` package in R [38].

The γ-statistic measures the distribution of branch lengths of a tree and is defined as [35]:

$$\gamma = \frac{1}{D}\left[\frac{1}{N_S - 2}\sum_{k=2}^{N_S-1} T(k) - T(N_S)/2\right] \tag{10}$$

with

$$T(k) = \sum_{j=2}^{k} j g_j; \tag{11}$$

$$D = T(N_S)/\sqrt{12(N_S - 2)} \tag{12}$$

where N_S is the number of leaves and g_k is the time interval between speciation events as represented by the nodes of the tree (see Fig. S4 in section III of the SI). The γ-statistic was calculated using the `gammaStat` function of the `ape` package in R [30].

The Sackin index measures the degree of imbalance, or asymmetry, of a tree [36, 37]. It is defined as

$$I_s = \sum_j d_j \tag{13}$$

in which d_j is the number of nodes to be traversed between each leaf j and the root, including the root [39]. The expected Sackin index under a pure birth process (the Yule model [40]) is

$$E(I_s(N_S)) = 2N_S \sum_{k=2}^{N_S} \frac{1}{k} \approx 2N_S \log N_S \tag{14}$$

where the approximation holds for N_S large [37]. Since the expected value of the Sackin index increases with the tree size, a normalized index is defined to compare trees of different sizes:

$$I_s^n = \frac{I_s(N_S) - E(I_s(N_S))}{N_S} \tag{15}$$

Here we used the normalized Sackin index to compare the phylogenies and calculated it using the `sackin` function of the `apTreeshape` package in R [41].

4. Results

We ran simulations of the speciation model described in section 3.1 with parameters $N = 1500$, $L = 100$, $B = 150$, $S = 5$, $G = 7$, $\mu = 0.001$, $D = 0.05$, $Q = 0.05$. We start with the results of a single simulation to show examples of phylogenies. Figure 4 shows the population after 1000 generations, with squares representing individuals and colors indicating the 36 species generated. Species form spatial clusters, a consequence of the small S value used the simulation.

The true phylogenetic tree of the population, generated using the SSEE algorithm, is shown in Fig. 5. Figure 5(a) shows the full phylogeny, which includes all speciation and extinction events. The large number of events seen near the root of the tree correspond mostly to unsuccessful or incomplete speciation events, in which a group of individuals momentarily splits in two species but quickly recombines into a single species due to mutations. We distinguish these events from *true extinctions*, which are characterized by the collapse of a long living species by a sharp decline in population size. This phenomenon is very common at the beginning of the speciation process in the model described in section 3.1. In Fig. 5(b),(c),(d) the full phylogeny was filtered in order to remove speciation reversals and keep only *true* extinction events. In the model, extinctions occur by stochastic fluctuations in the number of individuals of a species, which might become very small and go to zero. Figure 5(b) shows the phylogeny filtered by the criterion of population size at the moment of vanishing: species that disappear with more than 20 individuals were considered speciation reversals and removed from the tree. Figures 5(c) and (d) display the same phylogenies but filtered also by the criterion of persistence in time:

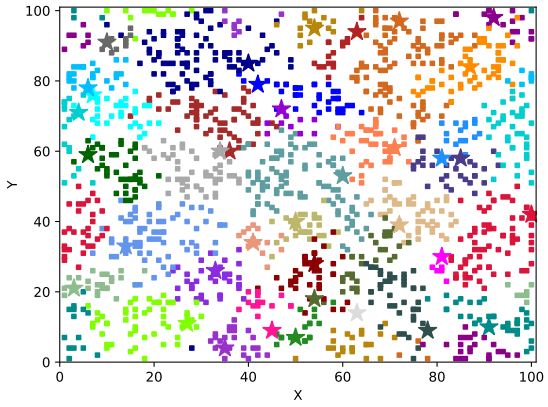

Figure 4: Spatial distribution of individuals from one simulation based on the model described in section 3.1. Individuals are represented by circles, and each color represents a different species. Stars indicate the individuals used to draw the phylogenies shown in figure 6.

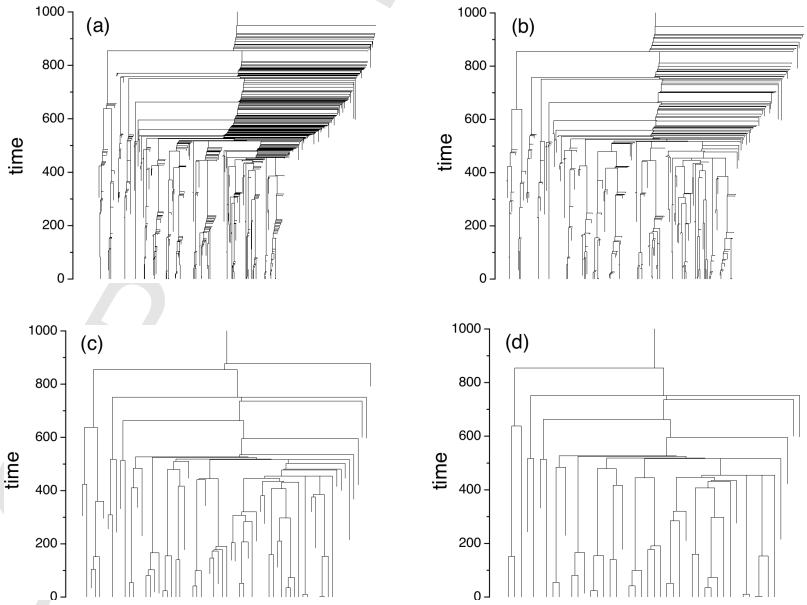

Figure 5: True phylogenies obtained with the SSEE method. (a) full phylogeny, including all speciation and extinction events; (b) filtered phylogeny, excluding branches (species) which had more than 20 individuals at the moment of extinction; (c) filtered phylogeny, excluding also branches that lasted less than 50 generations and (d) 100 generations.

branches of species that lasted less than 50 generations (c) or 100 generations (d) were also removed.

Phylogenies computed from the SSEE, MRCAT and genetic distance matrices are shown in Fig. 6. Panel (a) shows the true SSEE phylogeny, filtered to exhibit only the extant species. Panel (b) was obtained from the MRCAT algorithm, with one individual from each species being selected to represent the species. We showed in section II of the SI (Fig. S2) that the choice of the individual for constructing the phylogenetic tree with MRCAT can matter. However, the final structure of the tree will barely vary. Finally, panel (c) shows the phylogeny estimated from the genetic distance matrix of the same individuals used in Fig. 6(b) by the UPGMA clustering method. Differences in topology and branch lengths are qualitatively visible between these trees. Maternal and paternal genealogies obtained from the MRCAT algorithm are shown in Fig. S3 in the SI.

Statistical comparisons between phylogenies generated by the MRCAT algorithm and by the genetic distance methods (UPGMA, NJ and ME) in relation to the true phylogeny (SSEE) are shown in Fig. 7. The first line shows comparisons of topology (RF metric), branch length distribution (γ-statistic) and degree of imbalance (Sackin index) among phylogenies after 500 generations in 50 simulations. The second line shows the same comparisons after 1000 generations for the same 50 simulations. Colors represent the different methods utilized to generate the trees. In the RF scatterplots (Fig. 7(a)(b)) the coordinates of each point refer to the normalized topological distance between the tree calculated with the MRCAT matrix (y-axis) or by genetic distance matrix (x-axis) from the true phylogenies generated by the SSEE algorithm. Small values of RF indicate that phylogenies are closer to the true phylogeny (SSEE). The diagonal dotted line defines the condition in which the topology of the phylogenies (RF-value) was equal in trees generated by genealogical relationships (MRCAT trees) and that estimated by genetic distance (UPGMA, NJ and ME methods). The scatterplot for $T = 500$ (Fig. 7(a)) shows that phylogenies generated by MRCAT and genetic distance using UPGMA method (orange points) were similar in their RF-values, while trees estimated from NJ and ME methods (yellow and pink) had more different RF-values. For $T = 1000$ (Fig. 7(b)) all phylogenies estimated by genetic distance-based methods differ from those obtained by MRCAT. The density distribution of RF values shown above the scatterplots indicates that MRCAT is always closer to SSEE, especially for $T = 1000$.

Regarding the branch length distribution, the scatterplots (Fig. 7(c),(d)) show the difference between γ-values in SSEE phylogenies (y-axis) and MRCAT or genetic distance (UPGMA, NJ or ME) phylogenies (x-axis). The diagonal dotted line defines the condition in which the γ-values of trees generated by genealogical relationships (MRCAT trees) or by genetic distance (by UPGMA, NJ and ME methods) were equal to values of true phylogenies. We observe that for both times (Fig. 7(c),(d)) MRCAT trees had γ distributions closer to true phylogenies (SSEE) than all genetic distance-based trees, with a good match for $T = 1000$. Finally, the normalized Sackin index is presented in Fig.

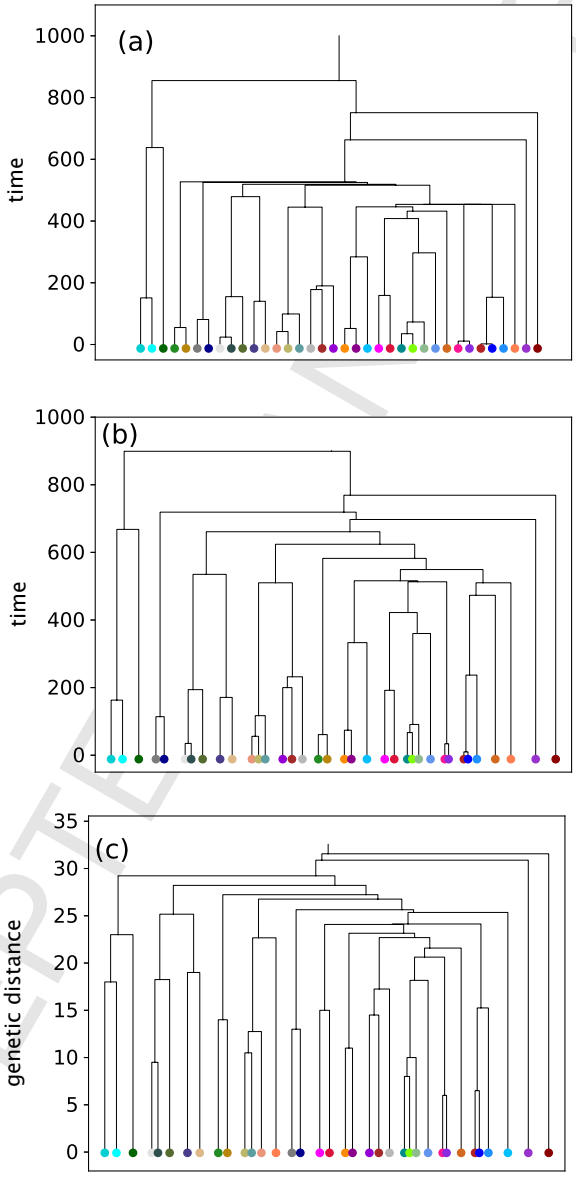

Figure 6: (a) Extant phylogeny obtained via SSEE (species are separated by one unit on x-axis); (b) via MRCAT; (c) via genetic distance matrix using UPGMA (neighbor species are separated by genetic distances). Colors correspond to species in Fig. 4.

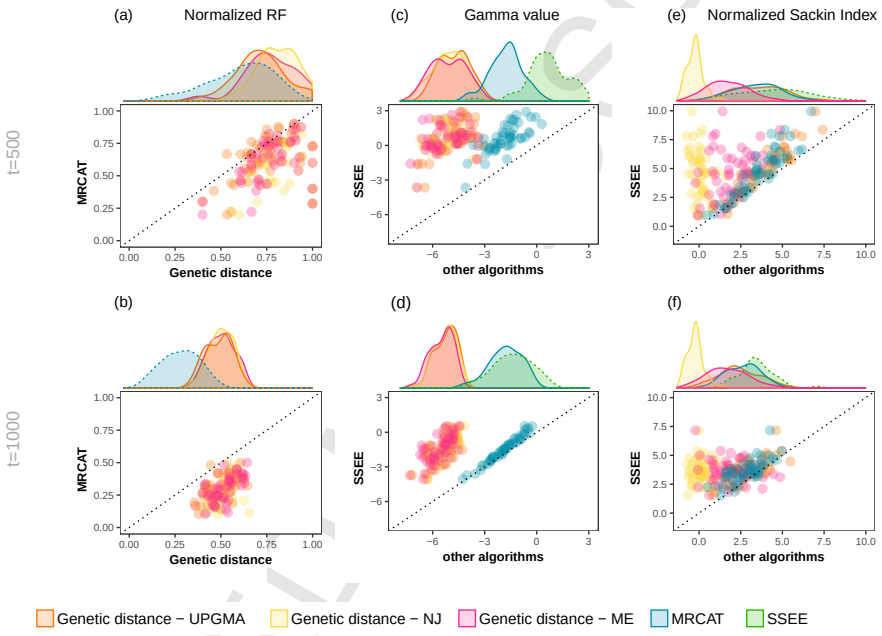

Figure 7: Comparisons among phylogenies generated by the algorithms proposed here (MRCAT and SSEE) and phylogenies estimated from genetic distance by UPGMA, NJ and ME methods. Lines exhibit the comparisons of RF, gamma and Sackin's metrics of 50 simulations at times 500 (first line) and 1000 (second line) generations. Colors represent the different methods utilized to generate the trees. (a) and (b): difference between RF-values of phylogenies obtained by MRCAT (y-axis) and by genetic distance-based methods (x-axis). Small values of RF indicate that phylogenies are closer to the true phylogeny (SSEE). (c) and (d): difference between branch length distributions (γ) of phylogenies generated by SSEE (y-axis, green distribution) and MRCAT algorithm (blue) or genetic distance-based methods (orange, yellow and pink) (x-axis). (e) and (f): the same as (c) and (d), but considering now the degree of imbalance (Sackin index). Distributions above all scatterplots illustrate qualitatively the differences in topology (a,b), branch length distribution (c,d) and degree of imbalance (e,f) of phylogenies generated from each algorithm or method in the 50 simulations.

(Fig. 7(e),(f)). The imbalance of MRCAT phylogenies was closer to the true phylogenies for $T = 500$ (Fig. 7(e)). On the other hand, for $T = 1000$ the imbalance was similar for MRCAT and all distance-based methods, except for the NJ. The NJ trees exhibited the most incorrect Sackin index (Fig. 7(e)(f)), possibly because NJ trees are not rooted, a necessary condition to compute this index. The rooting procedure chosen can be quite arbitrary, affecting the balance of the trees and consequently the Sackin index. The distributions above all scatterplots show qualitatively the differences in topology (Fig. 7(a),(b)), branch length distribution (Fig. 7(c),(d)) and degree of imbalance (Fig. 7(e),(f)) of phylogenies generated from each algorithm or method in the 50 simulations performed in each time ($t = 500$ or $t = 1000$).

5. Discussion

Understanding all the mechanisms that promote speciation is still an open problem in evolutionary biology [3, 42]. Even more challenging is to identify which of these mechanisms were important in a particular case. A large number of mathematical and computational models were developed in the past years to understand different speciation processes, such as neutral [43–46], sexual [47–49] and ecological selection [12, 50]. Models have also considered the role of geography in speciation, such as allopatric [51–54], parapatric [10, 55] and sympatric [12, 49, 56, 57]) scenarios. The results of models, however, can seldom be compared with real data [58, 59]. In these cases comparisons are often made in a macroecological scale, including qualitative species abundance and spatial distributions, species-area relationships and genetic or phenotypic distances [2, 6, 7, 14, 15]. Nevertheless, little attention has been given to the evolutionary history of individuals and species, neglecting the macroevolutionary scale underlying the speciation process [16, 21].

In this paper we have described two procedures to register the history of individuals (MRCAT) and species (SSEE) in individual-based models. With the ancestor-descendant relationships or speciation events saved in MRCAT and SSEE matrices we have constructed trees using a clustering algorithm. These trees have properties demonstrated in section I of Supporting Information. In the MRCAT algorithm, genealogies of individuals and phylogenies of species were obtained, whereas in the SSEE algorithm only phylogenies of species can be accessed. In the SSEE algorithm speciation events are precisely recorded and the resulting phylogenetic tree is the *true* tree of the community, whereas in the MRCAT algorithm the relations among species are recovered from genealogical relationships between individuals of each species. The MRCAT algorithm allows the construction of maternal, paternal and general lineages, the last being analogous to cases with hermaphroditic individuals. We have applied these algorithms to a spatially explicit IBM where individuals are separated into males and females and sexual reproduction is restricted by genetic difference below a threshold and by spatial proximity. We showed that maternal, paternal and general genealogies generated from the MRCAT algorithm are different even if the same individuals are chosen to draw the trees (Supporting Information,

section II). Maternal and paternal genealogies (Fig. S3(a),(b)) are different because they were obtained from different MRCAT matrices. In the first case, the MRCAT matrix contains the time to the most recent common *female* ancestor between each pair of individuals, while in the second case the MRCAT matrix has the time to the most recent common *male* ancestor between the same individuals, which lead to different ancestor times and genealogical relationships. In addition, for the general genealogy - taking the most recent common ancestor among females and males (*i.e.*, disregarding sex) - the resulting MRCAT matrix does not uniquely specify the genealogy (Fig. S3(c)). Regarding the phylogenetic trees, we showed that they may be different if obtained by MRCAT or SSEE algorithm (Fig. 6(a),(b), Fig. 7). As discussed in subsection 2.3, this mismatch happens because the time to the most recent common ancestor between individuals of different species is only an approximation to the speciation time, since speciation can happen several generations later (Fig. 3).

Structural properties of phylogenies, such as the Sackin index and the gamma distribution, obtained from SSEE and MRCAT trees were compared to values calculated in phylogenies estimated from the genetic distance between individuals of extant species by distance-based methods (UPGMA, NJ and ME). The aim of this comparison was to show that the validity of these methods commonly used in empirical studies, where the complete past history is inaccessible, can be assessed with the help of models. Differences in topology and branch length distribution measured by the RF metric and γ-statistic, respectively, revealed that MRCAT trees were closer to the true phylogenies (SSEE) than genetic distance-based trees. The difference between the results of these two methods possibly lies in back mutations that can happen in the genome of individuals, erasing the information needed to uncover the real history among species [60]. This phenomenon is more likely to happen at long times and for small genome size. Indeed, we observed that in 500 generations (Fig. 7(a)(c)) the phylogenies estimated from genetic distance were closer to the ones generated from MRCAT algorithm than in 1000 generations (Fig. 7(b)(d)), because in the first case the number of back mutations were probably smaller. Another factor that might explain the difference between genetic distance-based and true phylogenies is the sampling of only one individual to estimate the trees in the first case [61]. However, phylogenies generated with MRCAT algorithm also used only one individual per species - the same individuals used to compute genetic distance indeed - which suggests that this is not a very important factor (Fig. 7(a),(b),(c),(d)). The degree of imbalance showed a different picture, with less differences between MRCAT trees and genetic distance trees. Still, MRCAT trees were closer to the true phylogenies than the others. Trees estimated from genetic information in IBMs should be closer to the true phylogenies for larger genome sizes, where the probability of back mutations is smaller. Individual-based models with large or infinite genome sizes already available [26, 62] would provide good tests for measuring the accuracy of trees obtained by distance-based methods.

The better performance of MRCAT algorithm in recover the topology and balance of phylogenetic trees is not surprising, since matrices generated from

this algorithm hold the exact times to the most recent common ancestors. However, this type of exact information cannot be recovered from empirical data of contemporary samples. On the other hand, distance-based methods are commonly used for inference of phylogenetic trees from empirical data [61]. The advantage of these methods, especially the NJ method, is their computational efficiency. Indeed, cluster algorithms are faster than optimality criteria used in character-based methods, like maximum parsimony and maximum likelihood [61, 63]. Distance methods are particularly useful for analysis of data sets containing sequences with low levels of divergence [61]. However, methods based on genetic distances can perform poorly when the data set contains sequences with high levels of divergence due to greater sampling error in larger genetic distances. As most distance-based methods do not account for the high variances of large distance estimates, the inference of phylogenetic relationships could be impaired when these methods are employed [61]. In our model, trees generated from genetic distance methods were more different from the true trees (SSEE) than MRCAT phylogenies possibly because of high divergence among simulated genomes. This also could explain the high similarity in tree summary statistics among distance methods (Fig. 7). Moreover, the worst performance of NJ method in recover tree balance might be due to the lack of an explicit optimization criterion in the selection of taxon pairs in the original method proposed by Saitou and Nei [24] and utilized here [30, 63]. In addition, the choice of a substitution model to compute the pairwise distance between sequences might be important to determine the efficacy of distance methods [61]. Here we used the Hamming distance to calculate differences between pairs of sequences, but other methods could yield different results [64–67].

Modifications of the model to include *loci* not linked to the computation of genetic threshold would be important to understand how phylogenetic trees computed from these *loci* would differ from the ones computed here. Changing parameters values such as genome size and mutation rate could also affect tree estimations from distance-based methods and are a possible direction to future research. Nevertheless, the incorporation of algorithms that record the evolutionary history of individuals and species in an IBM context is an important step to help understanding the patterns left by specific speciation mechanisms at the macroevolutionary level.

6. Conclusions

The recent interest in the role of evolutionary history to explain the spatial patterns of abundance and species diversity calls for the incorporation of phylogenetic trees in the speciation modeling approach. Phylogenetic trees are essential tools to understand macroevolutionary patterns of diversity. They reveal how species are related to each other and the times between speciation events. Moreover, topological structure and branch length distribution also contain clues about processes originating a particular group of species. Previous works have already considered this problem for simpler models where each mutation corresponds directly to a new species [16]. Our study provides the

first general attempt to extend individual-based models by incorporating the branching process using the ancestor-descendant relationships between individuals and species. We believe this methodology will help predict and classify the macroevolutionary branching process, as well as the corresponding macroecological patterns (*e.g.*, species abundance distributions), resulting from different speciation models. The comparison of these results with empirical studies may clarify the role of different processes in generating the patterns observed in nature [4, 5]. Finally, the role of extinction in determining macroevolutionary patterns is an open field [19] which could be explored by using the full phylogenetic trees generated from the SSEE algorithm introduced here.

Acknowledgments

This research was supported by Sao Paulo Research Foundation (FAPESP), National Council for Scientific and Technological Development (CNPq), and Coordination for the Improvement of Higher Education Personnel (CAPES). We thank P. L. Costa for his constructive comments about the manuscript, and A. B. Martins and L. D. Fernandes, who provided expertise that greatly assisted the research, all contributing with their insights and comments about the research results.

Conflicts of interest: none.

[1] J. Coyne, H. Orr, Speciation, Sinauer Associates, Sunderland, MA, 2004.

[2] M. A. M. De Aguiar, M. Baranger, E. Baptestini, L. Kaufman, Y. Bar-Yam, Global patterns of speciation and diversity, Nature 460 (7253) (2009) 384–387.

[3] S. Gavrilets, Models of speciation: Where are we now?, Journal of heredity 105 (S1) (2014) 743–755.

[4] M. Turelli, N. H. Barton, J. A. Coyne, Theory and speciation, Trends in Ecology & Evolution 16 (7) (2001) 330–343.

[5] R. Field, B. A. Hawkins, H. V. Cornell, D. J. Currie, J. A. F. Diniz-Filho, J.-F. Guégan, D. M. Kaufman, J. T. Kerr, G. G. Mittelbach, T. Oberdorff, et al., Spatial species-richness gradients across scales: a meta-analysis, Journal of Biogeography 36 (1) (2009) 132–147.

[6] A. B. Martins, M. A. de Aguiar, Y. Bar-Yam, Evolution and stability of ring species, Proceedings of the National Academy of Sciences 110 (13) (2013) 5080–5084.

[7] F. May, A. Huth, T. Wiegand, Moving beyond abundance distributions: neutral theory and spatial patterns in a tropical forest, Proc. R. Soc. B 282 (1802) (2015) 20141657.

[8] M. Kopp, Speciation and the neutral theory of biodiversity: Modes of speciation affect patterns of biodiversity in neutral communities., BioEssays : news and reviews in molecular, cellular and developmental biology 32 (7) (2010) 564–70.

[9] S. P. Hubbell, The Unified Neutral Theory of Biodiversity and Biogeography, Princeton University Press, Princeton, NJ, 2001.

[10] S. Gavrilets, H. Li, M. D. Vose, Patterns of parapatric speciation, Evolution 54 (4) (2000) 1126–1134.

[11] U. Dieckmann, M. Doebeli, On the origin of species by sympatric speciation., Nature 400 (6742) (1999) 354–7.

[12] A. Rettelbach, M. Kopp, U. Dieckmann, J. Hermisson, Three modes of adaptive speciation in spatially structured populations, The American Naturalist 182 (6) (2013) E215–E234.

[13] S. Gavrilets, Perspective: models of speciation: what have we learned in 40 years?, Evolution 57 (10) (2003) 2197–2215.

[14] P. R. A. Campos, E. D. C. Neto, V. M. d. Oliveira, M. A. F. Gomes, Neutral communities in fragmented landscapes, Oikos 121 (11) (2012) 1737–1748.

[15] H. G. Martín, N. Goldenfeld, On the origin and robustness of power-law species–area relationships in ecology, Proceedings of the National Academy of Sciences 103 (27) (2006) 10310–10315.

[16] M. Manceau, A. Lambert, H. Morlon, Phylogenies support out-of-equilibrium models of biodiversity, Ecology letters 18 (4) (2015) 347–356.

[17] A. L. Pigot, A. B. Phillimore, I. P. Owens, C. D. L. Orme, The shape and temporal dynamics of phylogenetic trees arising from geographic speciation, Systematic biology 59 (6) (2010) 660–673.

[18] O. Hagen, K. Hartmann, M. Steel, T. Stadler, Age-dependent speciation can explain the shape of empirical phylogenies, Systematic Biology 64 (3) (2015) 432–440.

[19] T. B. Quental, C. R. Marshall, The molecular phylogenetic signature of clades in decline, PloS one 6 (10) (2011) e25780.

[20] T. J. Davies, A. P. Allen, L. Borda-de Água, J. Regetz, C. J. Melián, Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification, Evolution 65 (7) (2011) 1841–1850.

[21] J. Rosindell, L. J. Harmon, R. S. Etienne, Unifying ecology and macroevolution with individual-based theory, Ecology letters 18 (5) (2015) 472–482.

[22] D. L. DeAngelis, V. Grimm, Individual-based models in ecology after four decades, F1000prime reports 6.

[23] F. Murtagh, Complexities of hierarchic clustering algorithms: State of the art, Computational Statistics Quarterly 1 (2) (1984) 101–113.

[24] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees., Molecular biology and evolution 4 (4) (1987) 406–425.

[25] A. Rzhetsky, M. Nei, Theoretical foundation of the minimum-evolution method of phylogenetic inference., Molecular biology and evolution 10 (5) (1993) 1073–1095.

[26] P. G. Higgs, B. Derrida, Genetic distance and species formation in evolving populations, Journal of molecular evolution 35 (5) (1992) 454–465.

[27] E. M. Baptestini, M. A. de Aguiar, Y. Bar-Yam, The role of sex separation in neutral speciation, Theoretical ecology 6 (2) (2013) 213–223.

[28] R. W. Hamming, Error detecting and error correcting codes, Bell Labs Technical Journal 29 (2) (1950) 147–160.

[29] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2017).
URL `https://www.R-project.org/`

[30] E. Paradis, J. Claude, K. Strimmer, Ape: analyses of phylogenetics and evolution in r language, Bioinformatics 20 (2) (2004) 289–290.

[31] T. Britton, B. Oxelman, A. Vinnersten, K. Bremer, Phylogenetic dating with confidence intervals using mean path lengths, Molecular phylogenetics and evolution 24 (1) (2002) 58–65.

[32] L. J. Revell, S. A. Chamberlain, Rphylip: an r interface for phylip, Methods in Ecology and Evolution 5 (9) (2014) 976–981.

[33] J. Felsenstein, PHYLIP: Phylogenetic inference program, version 3.6 (2005).

[34] D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees, Mathematical biosciences 53 (1-2) (1981) 131–147.

[35] O. G. Pybus, P. H. Harvey, Testing macro–evolutionary models using incomplete molecular phylogenies, Proceedings of the Royal Society of London B: Biological Sciences 267 (1459) (2000) 2267–2272.

[36] M. Sackin, "good" and "bad" phenograms, Systematic Biology 21 (2) (1972) 225–226.

[37] M. G. Blum, O. François, On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited, Mathematical biosciences 195 (2) (2005) 141–153.

[38] K. P. Schliep, phangorn: phylogenetic analysis in r, Bioinformatics 27 (4) (2011) 592–593.

[39] B. L. Dearlove, S. D. Frost, Measuring asymmetry in time-stamped phylogenies, PLoS computational biology 11 (7) (2015) e1004312.

[40] G. U. Yule, A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs, Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character 213 (1925) 21–87.

[41] N. Bortolussi, E. Durand, M. Blum, O. Francois, apTreeshape: Analyses of Phylogenetic Treeshape, r package version 1.4-5 (2012).
URL `https://CRAN.R-project.org/package=apTreeshape`

[42] M. Kirkpatrick, V. Ravigné, Speciation by natural and sexual selection: models and experiments, The American Naturalist 159 (S3) (2002) S22–S35.

[43] G. A. Hoelzer, R. Drewes, J. Meier, R. Doursat, Isolation-by-distance and outbreeding depression are sufficient to drive parapatric speciation in the absence of environmental influences, PLoS Comput Biol 4 (7) (2008) e1000126.

[44] P. Desjardins-Proulx, D. Gravel, How likely is speciation in neutral ecology?, The American Naturalist 179 (1) (2011) 137–144.

[45] C. J. Melián, D. Alonso, S. Allesina, R. S. Condit, R. S. Etienne, Does sex speed up evolutionary rate and increase biodiversity?, PLoS Comput Biol 8 (3) (2012) e1002414.

[46] E. M. Baptestini, M. A. de Aguiar, Y. Bar-Yam, Conditions for neutral speciation via isolation by distance, Journal of theoretical biology 335 (2013) 51–56.

[47] G. S. van Doorn, P. Edelaar, F. J. Weissing, On the origin of species by natural and sexual selection, Science 326 (5960) (2009) 1704–1707.

[48] J. C. Uyeda, S. J. Arnold, P. A. Hohenlohe, L. S. Mead, Drift promotes speciation by sexual selection, Evolution 63 (3) (2009) 583–594.

[49] L. K. MGonigle, R. Mazzucco, S. P. Otto, U. Dieckmann, Sexual selection enables long-term coexistence despite ecological equivalence, Nature 484 (7395) (2012) 506–509.

[50] P. Nosil, Ecological speciation, Oxford University Press, 2012.

[51] J. L. Fierst, T. F. Hansen, Genetic architecture and postzygotic reproductive isolation: evolution of bateson–dobzhansky–muller incompatibilities in a polygenic model, Evolution 64 (3) (2010) 675–693.

[52] S. Gourbiere, J. Mallet, Are species real? the shape of the species boundary with exponential failure, reinforcement, and the missing snowball, Evolution 64 (1) (2010) 1–24.

[53] C. Fraïsse, J. Elderfield, J. Welch, The genetics of speciation: are complex incompatibilities easier to evolve?, Journal of evolutionary biology 27 (4) (2014) 688–699.

[54] R. Yamaguchi, Y. Iwasa, First passage time to allopatric speciation, Interface Focus 3 (6) (2013) 20130026.

[55] C. Bank, R. Bürger, J. Hermisson, The limits to parapatric speciation: Dobzhansky–muller incompatibilities in a continent–island model, Genetics 191 (3) (2012) 845–863.

[56] R. Bürger, K. A. Schneider, M. Willensdorfer, S. Otto, The conditions for speciation through intraspecific competition, Evolution 60 (11) (2006) 2185–2206.

[57] P. S. Pennings, M. Kopp, G. Meszéna, U. Dieckmann, J. Hermisson, An analytically tractable model for competitive speciation, The American Naturalist 171 (1) (2007) E44–E71.

[58] D. I. Bolnick, B. M. Fitzpatrick, Sympatric speciation: models and empirical evidence, Annu. Rev. Ecol. Evol. Syst. 38 (2007) 459–487.

[59] S. Gavrilets, J. B. Losos, Adaptive radiation: contrasting theory with data, Science 323 (5915) (2009) 732–737.

[60] J. Hein, M. Schierup, C. Wiuf, Gene genealogies, variation and evolution: a primer in coalescent theory, Oxford University Press, USA, 2004.

[61] Z. Yang, B. Rannala, Molecular phylogenetics: principles and practice, Nature Reviews Genetics 13 (5) (2012) 303.

[62] M. A. de Aguiar, Speciation in the derrida–higgs model with finite genomes and spatial populations, Journal of Physics A: Mathematical and Theoretical 50 (8) (2017) 085602.

[63] O. Gascuel, M. Steel, Neighbor-joining revealed, Molecular biology and evolution 23 (11) (2006) 1997–2000.

[64] T. H. Jukes, C. R. Cantor, et al., Evolution of protein molecules, Mammalian protein metabolism 3 (21) (1969) 132.

[65] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, Journal of molecular evolution 16 (2) (1980) 111–120.

[66] M. Hasegawa, H. Kishino, T.-a. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial dna, Journal of molecular evolution 22 (2) (1985) 160–174.

[67] Z. Yang, Estimating the pattern of nucleotide substitution, Journal of molecular evolution 39 (1) (1994) 105–111.