Controlled recovery of phylogenetic communities from an evolutionary model using a network approach

Arthur M. Y. R. Sousa[*]
Instituto de Física, Universidade Federal da Bahia, 40210-210, Salvador, Brazil
and Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, G3-52 4259 Nagatsuta-cho, Yokohama 226-8502, Japan

André P. Vieira[†] and Carmen P. C. Prado[‡]
Instituto de Física, Universidade de São Paulo, Caixa Postal 66318, 05314-970, São Paulo, Brazil

Roberto F. S. Andrade[§]
Instituto de Física, Universidade Federal da Bahia, 40210-210, Salvador, Brazil
(Received 8 November 2015; revised manuscript received 6 March 2016; published 27 April 2016)

This works reports the use of a complex network approach to produce a phylogenetic classification tree of a simple evolutionary model. This approach has already been used to treat proteomic data of actual extant organisms, but an investigation of its reliability to retrieve a traceable evolutionary history is missing. The used evolutionary model includes key ingredients for the emergence of groups of related organisms by differentiation through random mutations and population growth, but purposefully omits other realistic ingredients that are not strictly necessary to originate an evolutionary history. This choice causes the model to depend only on a small set of parameters, controlling the mutation probability and the population of different species. Our results indicate that for a set of parameter values, the phylogenetic classification produced by the used framework reproduces the actual evolutionary history with a very high average degree of accuracy. This includes parameter values where the species originated by the evolutionary dynamics have modular structures. In the more general context of community identification in complex networks, our model offers a simple setting for evaluating the effects, on the efficiency of community formation and identification, of the underlying dynamics generating the network itself.

DOI: 10.1103/PhysRevE.93.042317

I. INTRODUCTION

Deciphering the evolutionary history of living organisms is one of the major challenges of modern science [1]. The beginning of life on Earth can be tracked to some 3.8 to 4 billion years ago [2], and the first important attempts at reconstructing this process date from the 19th century, when Darwin's theory of evolution was proposed and incorporated as one of the cornerstones of science. Historically, theories and hypotheses about evolution are closely related to the setup of phylogenetic trees [2–4], which gather different extant species according to a suitable measure of their relative proximity, and to the study of fossils, which provided the first evidence of living species throughout evolutionary history, including many extinct ones.

The discovery of the DNA structure and, some decades later, the development of techniques to identify the genetic sequence of organisms yielded a huge amount of new information, with an impact on tasks, methods, and strategies of many fields of science [5,6]. Nevertheless, despite the large base of knowledge of genetic structures of extant and a few extinct organisms, phylogenies—and evolutionary reconstruction—are still essentially based on data of similarities between species alive nowadays [7]. There are, however, many limitations to this approach. The large variability in mutation rates of different genes and of similar genes in different groups of organisms, together with phenomena such as gene transposition and genome duplication, represent obstacles to a precise identification of phylogenetic trees and, as a consequence, to the reconstitution of evolutionary histories [8,9]. Moreover, since in general the exact evolutionary history of a group of species is unknown, there is no absolute way of checking the adequacy of those reconstructions, and different methods are validated by checking the mutual consistency of the corresponding reconstructions [10]. Reconstructed histories are also important in other fields such as linguistics (since language evolution follows some pattern of reproduction, mutation, and extinction) [11], but also in those cases validation is restricted to comparisons between different reconstruction methods.

In order to obtain a controllable comparison between evolutionary dynamics and phylogenetic classification, in the context of a recently proposed method [10,12], the present work considers a simple but well-defined computational evolutionary model and investigates the possibility of retrieving the corresponding phylogenetic communities from information on the set of species present at the final stage of the simulation process. The model contains the key ingredients of differentiation through random mutations and population growth, but we remark that other realistic ingredients were purposefully not included, seeking the minimum set of variables needed to mimic essential aspects of a phylogenetic tree. The idea is to provide an estimate of the maximum reliability of a phylogenetic reconstruction approach.

Our evolutionary model considers the microscopic evolutionary dynamics of a set of organisms characterized by a

[*]yamashita.a.ai@m.titech.ac.jp
[†]apvieira@if.usp.br
[‡]prado@if.usp.br
[§]randrade@ufba.br

genetic strand of binary bases. All organisms differentiate from a single common ancestor through a cumulative process in which random changes may occur in the strand. The model is therefore neutral, i.e., no explicit selection acts on any species. The phylogenetic classification is based on the identification of modules (or communities) in complex networks, whose structure is defined by the pairwise similarity index between the strands of the resulting organisms. We employ a community identification framework [10] that has already been used to investigate actual phylogenetic and evolutionary aspects based on molecular similarity in actual biological systems. We emphasize that this framework yields a phylogenetic classification in good agreement with other methods available in the literature [7], such as distance, maximum likelihood, maximum parsimony, and Bayesian methods, all of which, however, are subject to limitations due to the fact that we do not have full access to the actual evolutionary history.

On the other hand, since in our work we keep track of the evolutionary dynamics of the model, we are able to provide an absolute rather than relative estimate of the reliability of the chosen community-identification algorithm. Therefore, we are in a position to point out sources of errors between the predictions of the phylogenetic result and the actual evolutionary history. Of course this procedure can be used to test and establish benchmarks for any classification procedure and for any given evolutionary model. In the particular case of the framework of Ref. [10], given the consistency of the reconstructed phylogenies with those obtained from other methods, our estimated maximum reliability would be immediately applicable to those methods.

This paper is organized as follows: in the next section, we discuss our simple evolutionary model, which depends on four parameters; in Sec. III, we illustrate three possible evolutionary scenarios that are obtained by selecting different combinations of the model parameters. Section IV presents the major results of this work: the retrieved phylogenetic tree based on modularity analysis of complex networks. Concluding remarks and perspectives are mentioned in Sec. V.

II. EVOLUTIONARY MODEL

Here we present our simple evolutionary model that produces a set of species, their evolutionary history, and the genetic similarities between any pair of species. With such model, we can check the reliability of procedures for the inference of evolutionary history based on the similarity between species.

In our model, a species is identified by a strand of N units ("genes") that can take values 0 ("inactive gene") or 1 ("active gene").[1] Different gene sequences correspond to different species. Two species are said to be neighbor species if they differ by only one gene. For each species that appears in the model, we record information on the number of individuals belonging to that species.

The model is initialized with one species, which we call the species 0 (e.g., the species with all genes 0), with a single individual. In each time step T ("generation"), the number of individuals of an existing species i is updated according to a function $n_i(t - T_i)$, where T_i is the generation at which the species appeared, which requires that $n_i(t - T_i) \equiv 0$ for $t < T_i$. Then each individual of each existing species can make a transition to a random neighbor species ("mutation") with probability X, establishing an evolutionary link between those two species which will be registered in the evolutionary history. We stop the model at generation T_f, obtaining as outputs the set of generated species, their evolutionary history, and the similarity between each pair of species, as determined from the corresponding gene sequences. In Fig. 1, we depict the dynamics of the model. Notice that, contrary to work based on the fixed-population Wright-Fisher model (see, e.g., Refs. [9,13–15]), our focus here is not on the statistics of the distances between species or on features such as the time to the most recent common ancestor, but rather on both the genomes and the precise genealogy of the various species.

For simplicity, we constrain the mutation process so that a species can only mutate to one of its N neighbor species; that means that all of the links in the evolutionary history will connect neighbor species. Involution events, i.e., mutations to an already existing species, are not likely to happen for large enough N and small enough T_f, and are not explicitly prohibited in the model. Finally, the structure of the evolutionary history as well as the number of generated species are strongly dependent on the growth function $n(t)$ and the mutation probability X.

We use a dendrogram to picture the evolutionary history of the species. Representing successive generations in the horizontal axis, we can visualize the evolutionary history by observing the splitting of the branches, which indicates that mutations occur and new species appear from the species from which the branches stem. In order to construct the dendrogram, it is convenient to label the species in a specific order, so that the branches do not cross each other; we call this order the dendrogram numbering, which corresponds to the order of the branches at the right end of the dendrogram. Observe that the dendrogram numbering is different from the order according to which the species appear during the simulation, which we call the original numbering.

After the simulation ends, we build a similarity matrix $\mathcal{M}$, whose elements $\mathcal{M}_{ij}$ represent the similarity between species i and j, calculated as the ratio

$$\mathcal{M}_{ij} = \frac{\mathcal{G}_{ij}}{N}, \tag{1}$$

in which $\mathcal{G}_{ij}$ is the number of matching genes at corresponding positions in the gene sequences of species i and j, while N is the total number of genes. Notice that $\mathcal{M}$ is a symmetric matrix, since all species have the same total number of genes, appearing in the strand according to the same sequence. Furthermore, the diagonal elements of $\mathcal{M}$ are all equal to unity.

III. MODEL RESULTS

In this section, we show some results from the evolutionary model introduced in the previous section. The population

[1]The use of binary "genes" is reminiscent of the Penna model for biological aging [18,19].

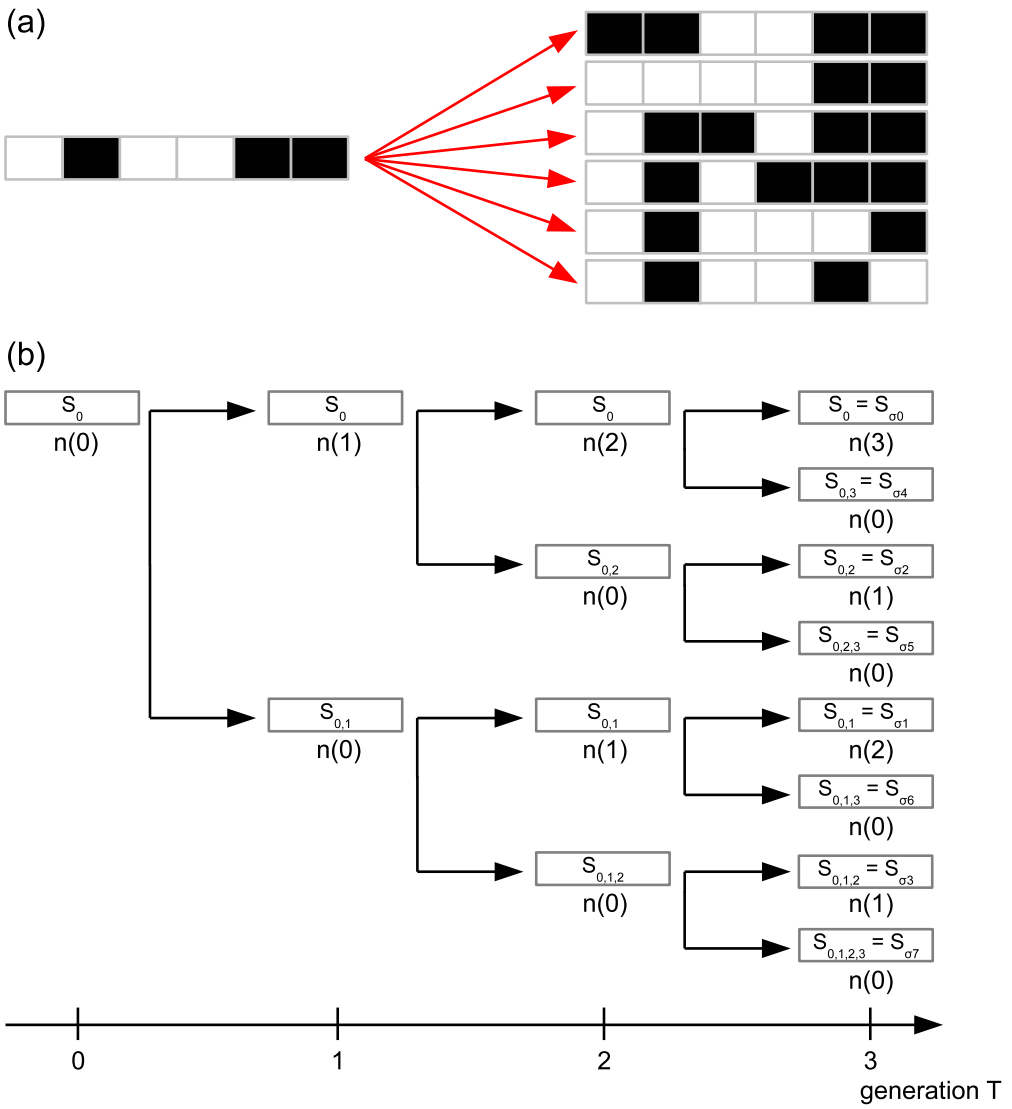

FIG. 1. Dynamics of the evolutionary model. (a) A species is represented by a strand of N genes (in the example, $N = 6$) that can be inactive (e.g., white) or active (black); in each step, this species can mutate with probability X to a neighbor species. (b) Scheme of the global dynamics of the model: the set of all species S is divided into subsets S_σ, where σ is a numeric sequence used to identify each subset. The numbers in the sequence indicate all generations at which the species belonging to that subset suffered a mutation. At $T = 0$, there is only one species and the subset containing it is identified by $\sigma_0 = 0$. At any $T \geqslant 1$, there are 2^{T-1} new subsets, labeled by 2^{T-1} new sequences σ_ℓ, $\ell = 2^{T-1}, 2^{T-1}+1, \ldots, 2^T - 1$. In particular, at $T = 1$, there is the σ_0 subset and a new one labeled $\sigma_1 = 0,1$. $\mathcal{N}_\sigma$ denotes the number of different species in S_σ, with the constraint $\mathcal{N}_{\sigma_0} = 1$. The value of $\mathcal{N}_\sigma$ depends on the random introduction of changes in the genome of the species ancestor. If no change occurs, $\mathcal{N}_\sigma \equiv 0$ for that particular sequence, as well as for all sequences resulting from adding new numbers to this sequence. $n_i(t - T_i)$ is the number of individuals of the species i at time t and, for any species i appearing at a time T_i, the condition $n_i(t < T_i) \equiv 0$ holds.

growth $n(t)$ is chosen as the logistic function,

$$n_i(t - T_i) = \frac{k e^{r(t-T_i)}}{k - 1 + e^{r(t-T_i)}}, \quad t \geqslant T_i$$
$$n_i(t - T_i) = 0, \quad t < T_i, \tag{2}$$

as it describes, in a simple way, the rate of population growth and the upper limit of the population size of a given species. Such features are in agreement with what is observed in actual organism populations. Here, r represents the growth rate and k is the carrying capacity, i.e., the maximum number of individuals of a given species. Notice that for a given value of k, $n(t)$ increases faster the larger the growth rate r. We remind the reader that if there were no upper limit on population size, the population of the set of older species would make them overwhelmingly dominate the speciation process. This tendency can be seen by looking at the results in Figs. 2–4, and comparing with the corresponding values r/k. In each of these figures, we draw the dendrogram obtained during the evolutionary history, the logistic growth function $n(t)$, and the gray tone (color in the online version) representation of the similarity matrix.

For the sake of definitiveness, we fixed the number of genes at $N = 10\,000$ and the probability of mutation at $X = 0.005$. We stop the simulation at the generation T_f, defined as

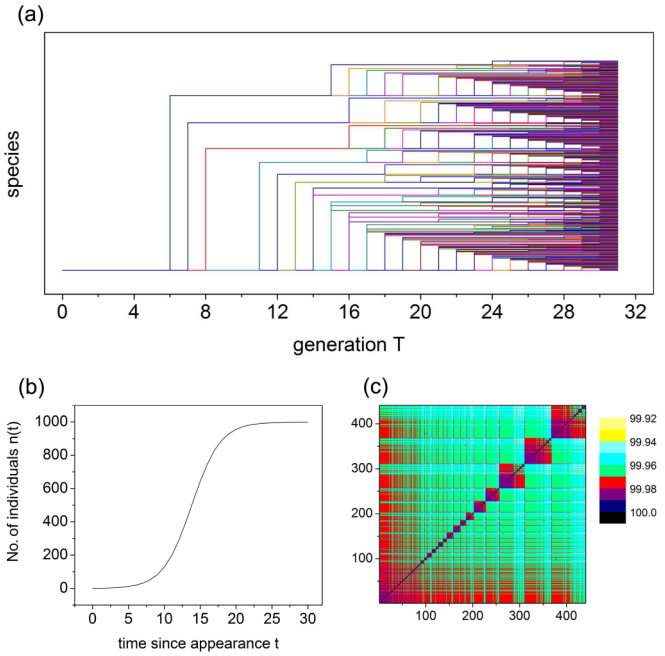

FIG. 2. (a) Dendrogram, (b) logistic growth $n(t)$, and (c) gray tone (color online) representation of the similarity matrix using the dendrogram numbering for the parameters $N = 10\,000$, $X = 0.005$, $T_f = 30$, $r = 0.5$, and $k = 1000$. For the sake of an easier identification of individual branches, gray tone (color) lines are used in (a), which have otherwise no special meaning. (c) The horizontal and vertical axes indicate the species, while the gray tone (color) bar indicates the genetic similarity between pairs of organisms, as defined by Eq. (1). The total number of species is 441. Notice the well-defined modular structures.

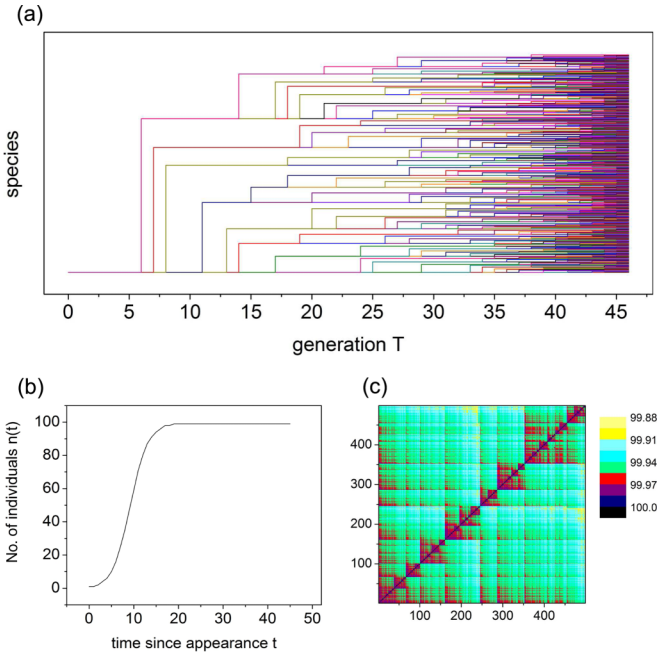

FIG. 3. (a) Dendrogram, (b) logistic growth $n(t)$, and (c) gray tone (color online) representation of the similarity matrix using the dendrogram numbering for the parameters $N = 10\,000$, $X = 0.005$, $T_f = 45$, $r = 0.5$, and $k = 100$. Horizontal and vertical axes, as well as gray tone (color) lines and pixels have the same meaning as in Fig. 2. The total number of species is 499. The modular structure is not so well isolated as in Fig. 2.

the generation in which the number of species exceeds a preestablished value $\bar{\mathcal{N}}$. In other words, the final set of species is composed of all species in the subsets S_σ (see the caption to Fig. 1), with $\sigma \leqslant \bar{\sigma}$, where $\bar{\sigma}$ is defined by the conditions

$$\sum_{\sigma=0}^{\bar{\sigma}} \mathcal{N}_\sigma \leqslant \bar{\mathcal{N}} \quad \text{and} \quad \sum_{\sigma=0}^{\bar{\sigma}+1} \mathcal{N}_\sigma > \bar{\mathcal{N}}. \tag{3}$$

All examples we discuss were obtained by setting $\bar{\mathcal{N}} = 500$. This maximal number of species is much smaller than the total number of possible species, 2^{10000}, reducing the probability of an involution event, and it is of the same order of the number of species used in some real phylogenetic analyses [10,12].

By varying the parameters r and k of the logistic growth function $n(t)$, we obtain different structures for the evolutionary history. The dependence on those parameters is intuitive since they determine the number of individuals that can mutate to new species. The carrying capacity k is related to the asymptotic maximum number of new species that can appear from a given species at a given generation; the growth rate r informs how fast $n(t)$ reaches the carrying capacity and thus how many species in a given generation are at their maximum number of individuals. Observing those aspects, we could distinguish three types of evolutionary history, as reflected in the respective dendrogram and similarity matrix.

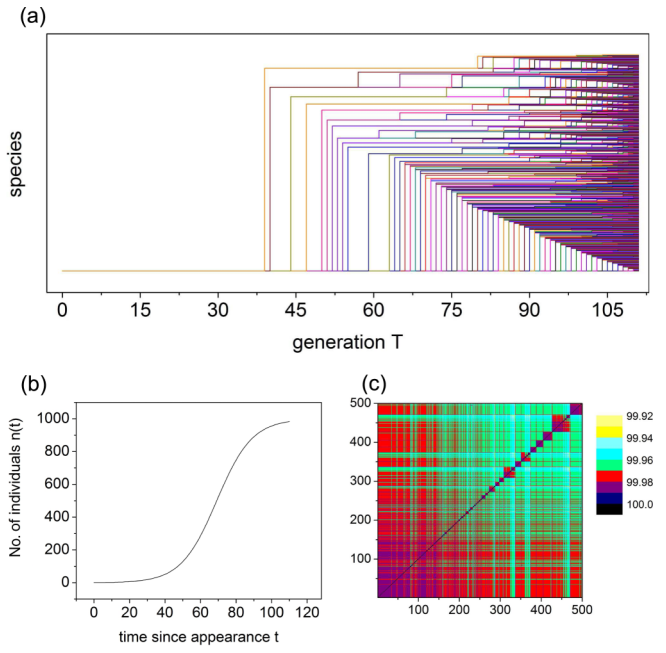

FIG. 4. (a) Dendrogram, (b) logistic growth $n(t)$, and (c) gray tone (color online) representation of the similarity matrix using the dendrogram numbering for the parameters $N = 10\,000$, $X = 0.005$, $T_f = 110$, $r = 0.1$, and $k = 1000$. Horizontal and vertical axes, as well as gray tone (color) lines and pixels have the same meaning as in Fig. 2. Total number of species is 500. Notice the large intertwined community comprising neighbors of the original species in the lower left corner.

The first type of structure is represented in Fig. 2. This case, for which $r = 0.5$ and $k = 1000$, allows the development of a well-defined modular structure, which can be observed in both the dendrogram and the similarity matrix. In the similarity matrix (using the dendrogram numbering), the species with high similarity are grouped together, resulting in the block diagonal structures of different sizes that indicate the existence of communities.

In the second type of structure, illustrated in Fig. 3, we keep the growth rate fixed at $r = 0.5$ but reduce the carrying capacity to $k = 100$. The modular structure is not well developed because, due to the smaller value of k when compared to the first case, at each generation only a small number of species appear, and almost all species reach the carrying capacity along the larger number of generations needed to reach the maximum number of allowed species. Although we can see some small block structures in the similarity matrix, they are not as well delimited as in the first case.

Finally, Fig. 4 shows a third kind of structure for the evolutionary history. Here, we keep $k = 1000$, the same value as for the first case, and reduce the growth rate to $r = 0.1$, therefore also slowing the rate at which new species appear. As a result, a large fraction of the species that appear up to the final considered generation T_f are neighbors of the original species, giving rise to a big module. The similarity matrix exhibits a large block structure corresponding to the species that are neighbors of the original one, and some other small blocks from species that are yet in the first stages of their growth.

We should stress that we explored only the case in which all species follow the same growth function $n(t)$—a logistic function with the same parameters r and k—and do not interact with each other. Also, our model does not involve the possibility of extinctions, a feature that can be introduced, for instance, by modifying $n(t)$.

IV. COMPARING EVOLUTIONARY HISTORY AND COMMUNITY-IDENTIFICATION RESULTS

Now we focus on the first type of evolutionary history, in which the modular structure is evident. Our aim is to check whether that evolutionary history can be recovered by using the framework of Refs. [10,12]. We stress that when comparing between the reconstructed phylogenies produced by different traditional methods, only a mutual consistency check is possible, as the precise evolutionary history is unknown. Here, on the other hand, we have access to the full simulated evolutionary history, thus allowing an absolute check on the reconstruction.

The first step involves using the similarity matrix $\mathcal{M}$ to define a set of networks, in which each node corresponds to a given species, while edges are drawn between each pair of vertices i and j for which $\mathcal{M}_{ij}$ is larger than a threshold λ. Here, the nodes are identified by the original numbering. The topology of the resulting network is strongly dependent on λ, and in particular we can define a distance $\delta(\lambda_1,\lambda_2)$ [16] between the networks characterized by the thresholds λ_1 and λ_2. As detailed in Refs. [10,12], this distance is derived from the elements of the neighborhood matrices of the networks.

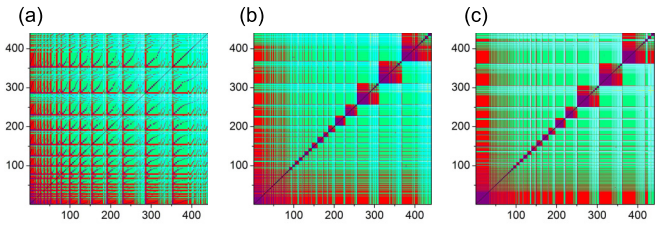

FIG. 5. Gray tone (color online) representation of the (a) similarity matrix based on the original numbering, (b) similarity matrix based on the dendrogram numbering, and (c) similarity matrix obtained from the NG numbering, for the model parameters used in Fig. 2. Horizontal and vertical axes, as well as the gray tone (color) code, used to indicate the genetic similarity between pairs of organisms, are the same as those used in Fig. 2.

Analyzing the behavior of $\delta(\lambda,\lambda+\Delta\lambda)$, i.e., the distance between networks obtained by values of λ differing by a small amount $\Delta\lambda$, there are peaks at all values of λ for which the topology of the corresponding networks is highly sensitive to small variations in the similarity threshold.

Previous work [10] has shown that the most pronounced peak in $\delta(\lambda,\lambda+\Delta\lambda)$ provides the optimal choice of λ for the identification of communities corresponding to phylogenetically related groups in real biological systems. For those real systems, the similarity matrix was calculated by analyzing enzymatic amino acid sequences which are not of

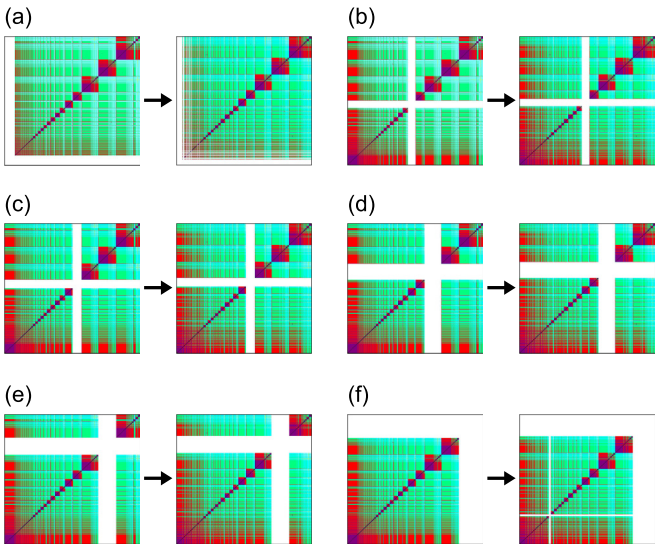

FIG. 6. Gray tone (color online) representation of NG and dendrogram similarity matrices for the model parameters used in Fig. 2. Comparison between the two representations is obtained by projecting the communities of the former onto the latter: (a) Community comprising species from 1 to 34 in the NG similarity matrix and the corresponding species in the dendrogram similarity matrix; (b) from 197 to 221; (c) from 222 to 250; (d) from 251 to 304; (e) from 305 to 361; and (f) from 362 to 441. Horizontal and vertical axes, as well as the gray tone (color) code, used to indicate the genetic similarity between pairs of organisms, are the same as those used in Fig. 2. For the sake of cleaner illustrations, tick labels have been removed from all panels.

TABLE I. Degree of correspondence between communities in the NG and the dendrogram (D) similarity matrix. The weighted average of the correspondence for the different communities is 93% for the entries corresponding to Fig. 2, 100% for Fig. 3, and 70% for Fig. 4.

Community	No. NG[a]	No. D[b]	No. NG/No. D
$T_f = 30$, $r = 0.5$,	$k = 1000$		
(a) 1–34	34	20	0.588
(b) 197–221	25	25	1.000
(c) 222–250	29	29	1.000
(d) 251–304	54	54	1.000
(e) 305–361	57	57	1.000
(f) 362–441	80	74	0.925
$T_f = 45$, $r = 0.5$,	$k = 100$		
(a) 1–37	37	37	1.000
(b) 38–66	29	29	1.000
(c) 67–100	34	34	1.000
(d) 101–141	41	41	1.000
(e) 142–201	60	60	1.000
(f) 202–267	66	66	1.000
(g) 268–352	85	85	1.000
(h) 353–499	147	147	1.000
$T_f = 110$, $r = 0.1$,	$k = 1000$		
(a) 1–102	102	33	0.324
(b) 344–365	22	22	1.000
(c) 366–391	26	23	0.885
(d) 392–419	28	28	1.000
(e) 420–457	38	32	0.842
(f) 458–500	43	42	0.977

[a]Number of species in the community according to the NG similarity matrix.
[b]Number of species in the community according to the dendrogram similarity matrix.

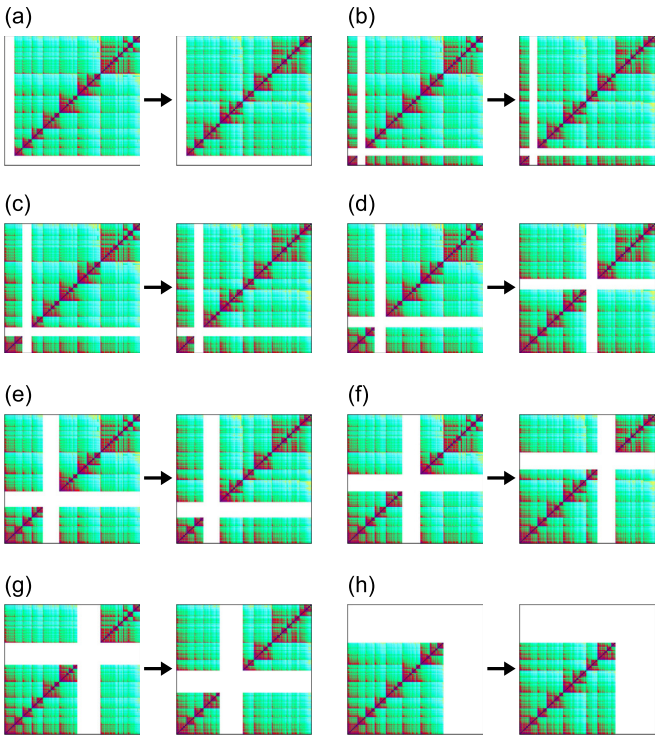

FIG. 7. Gray tone (color online) representation of NG and dendrogram similarity matrices for the model parameters used in Fig. 3. Comparison between the two representations is obtained by projecting the communities of the former onto the latter: (a) Community comprising species from 1 to 37 in the NG similarity matrix and the corresponding species in the dendrogram similarity matrix; (b) from 38 to 66; (c) from 67 to 100; (d) from 101 to 141; (e) from 142 to 201; (f) from 202 to 267; (g) from 268 to 352; and (h) from 353 to 499. Horizontal and vertical axes, as well as the gray tone (color) code, used to indicate the genetic similarity between pairs of organisms, are the same as those used in Fig. 3. For the sake of cleaner illustrations, tick labels have been removed from all panels.

equal length, and therefore also do not contain only equivalent subsequences. As a result, the optimal value of λ typically lies between 30% and 60%. For the present model, however, all "genetic" sequences have the same length, and equivalent "genes" all appear in the same order, so that, unsurprisingly, we obtained an optimal value of λ equal to 99.99 for all three choices of the model parameters indicated in the previous section.

The last step in the community-identification framework of Refs. [10,12] involves applying the Newman-Girvan (NG) algorithm, which is based on the successive elimination, from the optimal network, of edges having the largest betweenness coefficients [17]. This leads naturally to a different labeling of the edges, and in order to compare the communities predicted by the algorithm with those effectively produced by the simulations, we must establish a mapping between the NG and the dendrogram numberings.

Figure 5 shows the similarity matrices using the original, the dendrogram, and the NG numberings. The NG numbering was obtained by applying the NG algorithm to the optimal network at $\lambda = 0.9999$ using the original numbering, whose corresponding matrix does not present block structures. At this value of λ, the optimal network comprises only $m = 443$ connections out of the total of $M = 97\,020$ connections in a complete graph. These structures, however, are evident when the nodes are numbered according to both the dendrogram and NG numberings. On the other hand, Fig. 6 uses the mapping of the NG numbering onto the dendrogram numbering to depict the projection of the various communities predicted by the NG algorithm onto the dendrogram similarity matrix. Defining a community as a block structure in the similarity matrix with 20 or more species, we could identify six communities in both matrices by visual inspection. It is clear that to a high degree of accuracy (see Table I), the predicted communities do correspond to phylogenetically related groups of species.

Figures 7 and 8 show similar results for model parameters used in Figs. 4 and 5. Here, at the same value of λ, the corresponding values of the number of connections m in the optimal network are, respectively, $m = 500$ and $m = 507$. The results indicate that in the case of Fig. 4, all groups of organisms belong to the same communities in both matrices. For the parameters in Fig. 5, all smaller groups are accurately retrieved, while the organisms in the first well-defined group in the lower left corner of the NG similarity matrix are split into smaller groups in the dendrogram similarity matrix. The differences in the community patterns are explained by the parameter choice used in Fig. 4, which corresponds to the

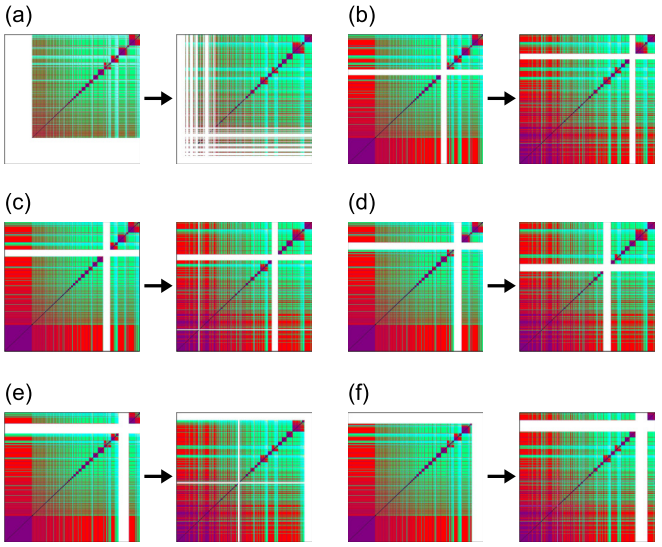

FIG. 8. Gray tone (color online) representation of NG and dendrogram similarity matrices for the model parameters used in Fig. 2. Comparison between the two representations is obtained by projecting the communities of the former onto the latter: (a) Community comprising species from 1 to 102 in the NG similarity matrix and the corresponding species in the dendrogram similarity matrix; (b) from 344 to 365; (c) from 366 to 391; (d) from 392 to 419; (e) from 420 to 457; and (f) from 458 to 500. Horizontal and vertical axes, as well as the gray tone (color) code, used to indicate the genetic similarity between pairs of organisms, are the same as those used in Fig. 4. For the sake of cleaner illustrations, tick labels have been removed from all panels.

smallest ratio $r/k = 10^{-4}$. This causes the population of the set of older species to overwhelmingly dominate the speciation process. As a consequence, they produced a huge number of different offsprings that differ in a very small amount from each other. The organisms enumerated according to the evolutionary history have a large intertwined structure when compared to that obtained by reconstructing the phylogeny from the final genes.

V. CONCLUSIONS AND PERSPECTIVES

We have introduced a simple model for the neutral evolution of species with bit-string genomes and a logistic population growth, aimed at producing a genealogy which allows absolute testing of phylogenetic reconstruction methods. In particular, our focus has been on a recently proposed method based on the distance between complex networks built from the similarities between the genomes of different species [10,12].

The results indicate that we achieved the aim of this work, namely, to show that it is possible to set up a simple evolutionary model for a two-purpose task: to follow the process of community formation and to recover, in a comparable way, the corresponding phylogenetic tree based solely on the "genetic" information extracted from the organisms in the last iteration of the growth process.

The results presented in the previous sections have shown that for a certain choice of parameters of the logistic population growth yielding a set of well-defined communities, e.g., for the first and second types of evolutionary history described in Sec. III, the approach of Refs. [10,12] is highly effective in reconstructing the phylogenetic tree from genetic data. This is an important result because it sets this efficiency in an absolute way since we now have access—for the computational model—to the full evolutionary history of the system.

Similar comparisons were performed for the second and third types of evolutionary histories, where the dendrogram numbering of the corresponding evolution histories leads to a somewhat blurred modular pattern. In the last case, our approach was less efficient to retrieve some of the community pattern generated by the evolutionary history. Nevertheless, other communities have been reliably recovered for both parameter sets. Despite the simplicity and the severe limitations of the present model, it is tempting to conclude that there are ranges of biological parameters (such as population sizes and the ratio between reproduction and mutation rates) which allow for a correct reconstruction of the phylogenetic relations between species, even in principle.

We are aware that the presence of interactions among the species, be it by competition leading to extinctions or by means of sexual reproduction, would necessarily insert new ingredients into the model. Given the very large number of possible mechanisms for speciation, following this path would go beyond the objectives of this work. Of course, it is expected that the introduction of such new rules' effects will lower the threshold value λ to obtain the optimal network, at which we perform the community identification. However, at this point we cannot offer any predictions as to the effects on the efficiency of community detection.

Finally, we remark that since our main goal was not to provide a realistic simulation of the evolution of species but rather to evaluate to what absolute extent current methods employed in phylogenetic reconstruction do manage to find correct evolutionary communities, our results seem already relevant. Of course, we would welcome new investigations examining more complete evolutionary models or spanning in more detail the parameter ranges for which the complex network approach has valid results.

ACKNOWLEDGMENTS

The authors are grateful to Professor A. Goés-Neto and Professor T. P. Lobão for helpful discussions. The authors also acknowledge the remarks of an anonymous referee during the review process. This work was partially supported by the Brazilian agencies FAPESB (Grant No. PNX 0006/2009), CNPq, INCT-SC, and NAP-FCx.

[1] A. Edwards and L. Cavalli-Sforza, in *Phenetic and Phylogenetic Classification*, edited by V. Heywood and J. McNeill (Systematics Association, London, 1964), Vol. 6, pp. 67–76.

[2] I. Fry, *The Emergence of Life on Earth: A Historical and Scientific Overview* (Rutgers University Press, New Brunswick, NJ, 2000).

[3] L. L. Cavalli-Sforza and A. Edwards, Am. J. Hum. Genet. **19**, 233 (1967).
[4] F. Ciccarelli, T. Doerks, C. von Mering, C. Creevey, B. Snel, and P. Bork, Science **311**, 1283 (2006).
[5] A. L. Barabási and Z. N. Oltvai, Nat. Rev. Genet. **5**, 101 (2004).
[6] G. Stoll and F. Naef, in *Algorithms and Computational Methods for Biochemical and Evolutionary Networks*, edited by M.-F. Sagot and K. S. Guimarães (College Publications, London, 2005), pp. 115–116.
[7] J. Felsenstein, *Inferring Phylogenies* (Sinauer Associates, Sunderland, MA, 2003).
[8] B. Mirkin, T. Fenner, M. Galperin, and E. Koonin, BMC Evol. Biol. **3**, 2 (2003).
[9] J. Sirén, W. Hanage, and J. Corander, Mol. Biol. Evol. **30**, 457 (2013).
[10] R. F. S. Andrade, I. C. Rocha-Neto, L. B. L. Santos, C. N. de Santana, M. V. C. Diniz, T. P. Lobão, A. Goés-Neto, S. T. R. Pinho, and C. N. El-Hani, PLoS Comput. Biol. **7**, e1001131 (2011).
[11] F. Petroni and M. Serva, J. Stat. Mech. (2008) P08012.
[12] A. Góes-Neto, M. V. C. Diniz, L. B. L. Santos, S. T. Pinho, J. G. Miranda, T. P. Lobão, E. P. Borges, C. N. El-Hani, and R. F. Andrade, Biosystems **101**, 59 (2010).
[13] M. Serva, J. Stat. Mech. (2005) P07011.
[14] D. Simon and B. Derrida, J. Stat. Mech. (2006) P05002.
[15] E. Brunet and B. Derrida, J. Stat. Mech. (2013) P01006.
[16] R. F. S. Andrade, J. G. V. Miranda, S. T. R. Pinho, and T. P. Lobão, Phys. Lett. A **372**, 5265 (2008).
[17] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[18] T. J. P. Penna, J. Stat. Phys. **78**, 1629 (1995).
[19] D. Stauffer, Bioinformat. Biol. Insights **1**, 91 (2007).