Eur. Phys. J. B **70**, 583–592 (2009)
DOI: 10.1140/epjb/e2009-00254-8

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Topological properties of phylogenetic trees in evolutionary models

M. Stich[a] and S.C. Manrubia

Centro de Astrobiología (CSIC-INTA), Ctra. de Ajalvir km. 4, 28850 Torrejón de Ardoz (Madrid), Spain

Received 16 April 2009
Published online 21 July 2009 –

Abstract. The extent to which evolutionary processes affect the shape of phylogenetic trees is an important open question. Analyses of small trees seem to detect non-trivial asymmetries which are usually ascribed to the presence of correlations in speciation rates. Many models used to construct phylogenetic trees have an algorithmic nature and are rarely biologically grounded. In this article, we analyze the topological properties of phylogenetic trees generated by different evolutionary models (populations of RNA sequences and a simple model with inheritance and mutation) and compare them with the trees produced by known uncorrelated models as the backward coalescent, paying special attention to large trees. Our results demonstrate that evolutionary parameters as mutation rate or selection pressure have a weak influence on the scaling behavior of the trees, while the size of phylogenies strongly affects measured scaling exponents. Within statistical errors, the topological properties of phylogenies generated by evolutionary models are compatible with those measured in balanced, uncorrelated trees.

PACS. 87.23.Kg Dynamics of evolution – 89.75.Hc Networks and genealogical trees

1 Introduction

Ever since the first observations on the diversity of living beings, there has been an interest in classifying them according to their similarities. As early as in the mid XV century, taxonomy jumped from folk inventories to global classification. By the end of the XVIII century, the taxonomic classification included about ten thousand species of plants and more than thousand different genera. The next level in the taxonomy, that of families, was also incorporated towards the end of that same century [1]. Still, the idea of a common origin for living beings was absent from that classification. It was only through the onset of an evolutionary theory, and especially after the publication of *The Origin* by Charles Darwin [2], that the nowadays iconic image of a tree of life relating extant organisms to extinct common ancestors began to take form.

It was soon observed that most taxonomic groups are species-poor, and only a few are composed of many species, this pattern repeating as one climbs up taxonomic levels [3]. The resulting hierarchical classification could, in the light of evolution, be viewed as a branching process in time, thus completely changing the interpretation and meaning of the data. The first model aimed at representing the common origin of species and their uneven distribution within the tree was that of Yule [4], which already yielded a remarkable agreement with empirical data. Yule's model is a neutral model of evolution that starts with a single species in the tree. The probability that a species splits into two is uniform in the tree and does not depend on time. The statistical properties of the genealogy so constructed are identical to those of the equal-rates Markov (ERM) model and of neutral coalescent models of phylogenetic trees [5].

The model proposed by Yule is a first instance of assimilating taxonomy to phylogeny. Actually, whether taxonomic classification is consistent with the actual phylogeny of species is a non-trivial question: while the former results from a largely artificial division, the latter explicitly follows the evolutionary history of a clade, and contains no visible division into groups. The robustness of statistical patterns in taxonomy, as obtained from different groups of animals and plants and at different taxonomic levels, seems to support the hypothesis that statistical properties of taxonomy do not depend on the details of the classification and contain reliable information about the patterns of biological diversity [6]. Still, it remains to be proved that the properties of taxonomy at the species level is equivalent to the statistical properties of higher-order taxa [7], although analyses of the topology of large trees seem to support that mechanisms driving biological diversification are independent of the taxonomic level [8].

Phylogenetic trees are nowadays routinely reconstructed by means of genomic data [9]. Molecular information on extant organisms (parts of genomes, single genes, mitochondrial RNA, proteins or even metabolic networks)

[a] e-mail: stichm@inta.es

can be used to determine an evolutionary distance between each pair of species. Different methods permit to reconstruct a tree where species sit at the tips of the tree, each associated with a branch. Two branches merge in a node, which stands for the common ancestor. The length of the branches, once calibrated, conveys information on the times in the past where the splitting occurred. Despite all the major advances in the reconstruction techniques [10], it is important to keep in mind that we cannot access the real phylogenetic tree containing the precise historical relationships between species.

Most phylogenetic trees available comprise several to hundred species. Trees with less than ten species are very abundant. The evolutionary time scale they reflect is usually relatively short, and the species in those trees are typically closely related. Small trees have been the focus of many studies dealing with their shape and its meaning in the evolutionary process [11]. Trees of medium size, between ten and hundred species, are also frequent. They may yield more reliable statistical measures and span larger time scales. Trees with more than hundred species, of which there are only a few, are considered large trees [5,12].

An often used measure to quantify to which extent branching probability deviates from homogeneity and affects the shape of phylogenetic trees is tree imbalance [5,11,13,14]. In a completely balanced tree each pair of daughter branches splits with the same probability; in a completely imbalanced tree only the left (or right) daughter branch splits. Natural trees or any tree generated through a model with a certain random component will be in-between those extremes. In particular, for small trees, a random process involved or an imprecise reconstruction can produce a considerable degree of imbalance. Only large enough trees can demonstrate whether imbalance is intrinsic to the evolutionary process or a finite-size effect. From a biological viewpoint, a balanced phylogenetic tree indicates that clades are equally likely to speciate. An unbalanced tree, on the other hand, reveals that species keep memory of the ability of their ancestors to diversify. In other words, if a species stems from a highly radiating group, it will be more prone to speciate in the future. This possibility has received some support from empirical analyses, though the inheritance of proneness to speciate is local in time [15]. There are additional evolutionary reasons that can induce a change in the speciation rate and thus affect tree imbalance [13]: refractory periods after speciation, adaptive radiations, selective extinctions, or fluctuations in the environment causing differences in the selection pressures. However, this memory of past success or of contingent evolutionary events should be observed only up to certain time scales, since there is a second process causing that memory to wear off: mutation. Mutations and adaptation to changing environments have the effect of erasing the memory of past success. There is evidence that correlations between the properties of distant species should decay exponentially fast due to mutation, even in the presence of persistent inheritance. This has been calculated for phylogenetic models of neutral genotypes [16]. In a very different framework, it has been observed how mutation erases spatial correlations exponentially fast in non-neutral models for competition between phenotypes [17].

Tree imbalance is a topological measure of tree shape that does not take into account branch length – i.e. evolutionary time – , and can be quantified in different ways [13]. The use of these measures goes beyond the phylogenetic context. For example, they have been used to compare between random trees and trees generated from spin-glass energy landscapes [18]. There, it has been shown that tree asymmetry of spin-model trees increases with the size of the tree. In this work, we use as topological measure the relation between the subtree size and the cumulative branch size, which have served to quantify the branching properties of transportation networks [19], food-webs [20,21], and phylogenetic trees [8,14], among others. In particular the latter works, where the analyzed data came from the most exhaustive database available, TreeBASE [12], and included trees with up to 600 species, served as motivation for our study.

The main aim of this work is to quantitatively study certain topological properties of phylogenetic trees generated by different models of evolving populations. This means that the nodes of the trees created by such models actually represent individuals (e.g. RNA molecules) rather than species. We then analyze how the topological quantitites depend on significant parameters as the mutation rate, the selection pressure, and the size of the tree.

The article is structured as follows: we begin by reviewing the topological properties of simple models statistically equivalent to the ERM (such as the backward coalescent) and analytically derive the relevant quantitites for simple trees, being of particular importance the case of completely balanced trees (Sect. 2). These first results are useful to compare with the topological characteristics of more sophisticated models. The core of our contribution is the analysis of different evolutionary population dynamics models where the ability of an ancestor to produce offspring is inherited and mutated with a variable probability (Sects. 3 and 4). This approach differs from models that consist of algorithms specifying how to construct a tree but lack any biological interpretation of its rules [7,14], and also from biologically motivated models including inheritance (i.e. memory) and mutation where, however, no selection is acting [16,22]. For a Moran's model, the effect of selection on the topology of small genealogical trees has been studied using alternative measures of tree imbalance [23].

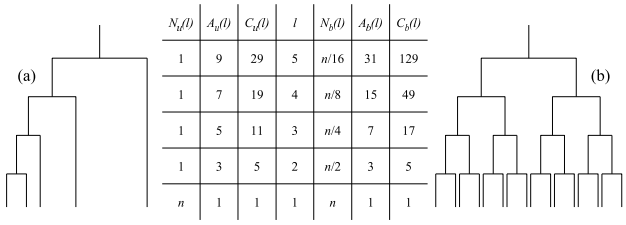
In Section 3, we use an explicit model of molecular evolution considering RNA sequences which are folded into their minimum free energy secondary structure. The distance from each folded sequence to a target secondary structure determines its ability to replicate. In Section 4, we introduce a simple model of a replicating population, where each element produces offspring according to its fitness. Mutation may increase or decrease the fitness of the individuals of the daughter generation. These simple models allow us to study very large systems and show

that, for small trees, the topological properties differ from the asymptotic properties of random models. The results for the models investigated in Sections 2–4 are compared in Section 5. There, we show that for large enough system size the topologies of all trees, i.e. also of the trees obtained from the explicit evolutionary models with selection, seem to be compatible with trees obtained from the class of ERM. The article is closed with a discussion of the results (Sect. 6).

2 Topology of simple trees

Two useful quantities to evaluate the topology of trees are the subtree size and the cumulative branch size. For each node i in the tree, the subtree size A_i is defined as the number of subtaxa diversifying from node i, including itself. The cumulative branch size C_i is defined as $C_i = \sum_j A_j$, where the sum runs over all nodes j diversifying from i, including itself. For a given tree, the probability distributions of A and C may display power-law tails, $P(A) \sim A^{-\alpha}$ and $P(C) \sim C^{-\gamma}$. Whenever there is a one-to-one relationship between A and C values, as in the cases we are going to discuss, it is of the scaling type, $C \sim A^{\eta}$, with $\eta = (1-\alpha)/(1-\gamma)$. The values of the exponents characterize the degree of imbalance of a tree. It has been shown that completely balanced trees are asymptotically described by $\alpha = 2$, $\gamma = 2$, and $\eta = 1$ – with a relevant logarithmic correction in $P(C)$ that we rederive below. The exponents characterizing completely unbalanced trees are $\alpha = 0$, $\gamma = 1/2$, and $\eta = 2$. An interesting example deviating from these extreme behaviors is the case of efficient transportation networks, whose topology is described by an exponent $\eta = 3/2$ that results from an optimization principle [19]. The scaling of food webs was first reported to yield a value for η between 1.13 and 1.16 [20]. Later, however, it was convincingly argued that the previous non-trivial value was a spurious result due to food webs having only a few trophic levels, and the exponent was corrected to $\eta = 1$ [21]. This is a first word of caution towards the meaning of topological quantities derived from small systems. Finally, critical branching trees [24] display $\alpha = 3/2$, and all supercritical branching trees follow $\alpha = 2$ [25]. This latter class is shared by the ERM, by Yule's model [4] and also by the coalescent [5]. This means that asymptotically the scaling of these models is characterized by the exponent $\eta = 1$, coinciding with the case of completely balanced trees. In Sections 4 and 5, we will present simulations of the coalescent model for comparison with the models introduced below.

In the remaining of this section we derive the probability distributions for completely balanced trees and completely unbalanced trees, and pay particular attention to the non-trivial logarithmic corrections: these cause a continuous bending of the distribution $P(C)$ that, as will be shown in forthcoming sections, may lead to estimations of the exponent η in small systems remarkably far from its asymptotic value.



$N_u(l)$	$A_u(l)$	$C_u(l)$	l	$N_b(l)$	$A_b(l)$	$C_b(l)$
1	9	29	5	$n/16$	31	129
1	7	19	4	$n/8$	15	49
1	5	11	3	$n/4$	7	17
1	3	5	2	$n/2$	3	5
n	1	1	1	n	1	1

Fig. 1. Simple trees and quantities characterizing their topology. (a) Completely unbalanced tree. (b) Completely balanced tree. Each tree starts with n leaves (tips). The table shows the level l in the tree, the number of nodes $N_{u/b}(l)$ for the unbalanced/balanced case and the corresponding values for the subtree size $A_{u/b}(l)$ and the cumulative branch size $C_{u/b}(l)$.

In the completely balanced and completely unbalanced trees, A_i and C_i only take a limited possible number of integer values (see Fig. 1). In order to quantitatively compare these results with other examples in the literature, we will assume a continuum approximation (as in [25]) to estimate the probability density distributions $P(A)$ and $P(C)$. We first calculate the number of nodes with each value of A and C and subsequently normalize dividing by the corresponding interval, ΔA and ΔC, between actually represented values. In this section, n denotes the number of tree tips.

We begin with the completely balanced tree. Let us call $N_b(l)$ the number of nodes at level l, and $A_b(l)$ and $C_b(l)$ the value of the branch size and the cumulative branch size, respectively, at that level. The interval lengths separating two consecutive values are $\Delta A_b = A_b(l+1) - A_b(l)$ and $\Delta C_b = C_b(l+1) - C_b(l)$. We apply $C_b(l) = 2C_b(l-1) + A_b(l)$, with the condition $C_b(1) = 1$ to solve the recursion for $C_b(l)$. Then, $N_b(l) = n/2^{l-1}$, $A_b(l) = 2^l - 1$, $C_b(l) = 1 + 2^l(l-1)$, $\Delta A_b = 2^l$, and $\Delta C_b = (l+1)2^l$.

To obtain expressions in terms of A_b and C_b, the parametric variable l is eliminated and we transform $N_b(l)$ to probability distributions $P_b(A)$ and $P_b(C)$ by dividing through the intervals ΔA_b and ΔC_b, respectively. This yields

$$P_b(A) = \frac{2n}{(A+1)^2}. \tag{1}$$

The parametric solution for $P_b(C(l))$ reads

$$P_b(C(l)) = \frac{n}{2^{2l-1}(l+1)}, \tag{2}$$

with

$$l = \frac{W(z)}{\ln 2} + 1, \tag{3}$$

where $z = (C-1)\ln 2/2$, and with $W(z)$ denoting the Lambert W-function, defined as the function satisfying $z = W(z)e^{W(z)}$. Thus, the distribution $P(C)$ cannot be obtained in an explicit analytical form. The Lambert W-function admits the following asymptotic expansion for $z \geq 3$ [26],

$$W(z) = \ln z - \ln\ln z + \frac{\ln\ln z}{\ln z} + O\left[\left(\frac{\ln\ln z}{\ln z}\right)\right]^2, \tag{4}$$

which substituted to first order in the expression for $P_b(C)$ eventually yields a functional form for large C,

$$P_b(C) \simeq \frac{n}{C^2 \ln 2(\ln(C \ln 2))}. \quad (5)$$

Corrections with arbitrary precision can be obtained through the use of additional terms in the series expansion of $W(z)$.

Finally, the function $C_b(A)$ has the exact form

$$C_b(A) = \frac{(A+1)\ln(A+1)}{\ln 2} - A, \quad (6)$$

thus presenting the well-known $A \ln A$ behavior that asymptotically yields the scaling $\eta = 1$, although it is important to keep in mind the logarithmic correction. Note that the previous derivation can be understood as a mean-field approximation to the topological properties of a tree whose nodes produce, on average, two branches per generation. The calculation can be further generalized to a tree whose branches split into k new branches on average at every level l. Taking now $C_b(l) = kC_b(l-1) + A_b(l)$, $C_b(1) = 1$, the relationship between C and A is

$$C_b^k(A) = \frac{1 + [A(k-1)+1]\left[\frac{\ln[A(k-1)+1]}{\ln k} - 1\right]}{(k-1)^2}. \quad (7)$$

Note that k does not change the functional form of $C(A)$: the scaling exponent remains unchanged and only the coefficients are modified.

For completely unbalanced trees, the relevant parametric quantities read $N_u(l) = 1$ if $l \neq 1$, $N_u(1) = n$, $A_u(l) = 2l - 1$, $C_u(l) = l^2 + l - 1$, $\Delta A_u = 2$, $\Delta C_u = 2(l+1)$, so

$$P_u(A) \propto \frac{1}{2} = \text{const.}, \quad (8)$$

$$P_u(C) \propto \frac{1}{2}\left(\frac{1}{C}\right)^{1/2} - \frac{5}{16}\left(\frac{1}{C}\right)^{3/2} + \frac{75}{256}\left(\frac{1}{C}\right)^{5/2} + O\left(\frac{1}{C}\right)^{7/2}, \quad (9)$$

and

$$C_u(A) = \frac{A(A+3)}{2}, \quad (10)$$

hence displaying a power-law scaling with exponent $\eta = 2$. It is interesting that the completely unbalanced tree presents a pure power law in the limit of large tree sizes, contrary to what occurs with the completely balanced tree.

3 Evolution of RNA populations

RNA is considered to be an appropriate model for studying evolution of populations [27,28]. Each RNA sequence can be mapped to a folded secondary structure of minimum free energy. The mapping between sequence and structure is degenerated (many sequences fold into the same secondary structure), which constitutes one of the most interesting properties of this model. RNA folding provides an explicit separation between genotype (represented by sequences, upon which mutations act) and phenotype (secondary structures, upon which selection acts). In this work, we use the evolution of populations of RNA sequences to study phylogeny while in previous work we focused on other aspects of the evolutionary process [29]. Since we have access to the exact genealogy of the population, we do not need to resort to reconstructed trees and hence eliminate one of the possible sources of error in quantifying phylogenetic properties. Further, the model is biologically grounded and has an explicit evolutionary mechanism. In this way, it overcomes one common criticism raised against some phylogenetic models, i.e. the absence of an appropriate measure of biological fitness [5].

3.1 Evolutionary algorithm

Our model system consists of a population of N replicating RNA sequences of constant length, subjected to point mutations and selection. The algorithm sketched below is described in more detail in reference [29].

Population sizes and sequence length are kept constant during simulations. Every molecule of the population is initialized with a random sequence of the four type of nucleotides A, C, G, and U. Every time that a molecule replicates, each of its nucleotides has a probability μ to be randomly replaced by another (or the same) type of nucleotide. This is how mutation is implemented. At each generation, the sequences are folded into secondary structures with help of the Vienna RNA package [30], version 1.5, used with the current standard parameter set. We define a target secondary structure which represents in a simple way optimal performance in the given environment. The target structure considered in this study corresponds to a biologically relevant RNA structure consisting of 165 nucleotides, the 165 *thiM* molecule, a riboswitch identified in the bacterium *Escherichia coli* [31]. For the aim of this study, choosing a particular target structure is not of crucial importance. We expect our results to hold qualitatively for other secondary structures. We assume that the probability that a sequence replicates is larger the more similar is its secondary structure to the target structure. In this way, sequences having structures similar to the target structure become more abundant. Eventually, for mutation rates below a critical mutation threshold, the population adapts to the environment, that is, it maintains a finite fraction ρ of sequences folding into the target structure. To quantify the similarity between a secondary structure of a given sequence and the target structure, we use the base-pair distance as implemented in the Vienna package [30]. The base-pair distance between two secondary structures is given by the number of base pairs that have to be opened and closed to transform one structure into the other. The probability $p(d_i)$ that sequence i replicates is given by

$$p(d_i) = Z^{-1} \exp(-\beta d_i / d), \quad (11)$$

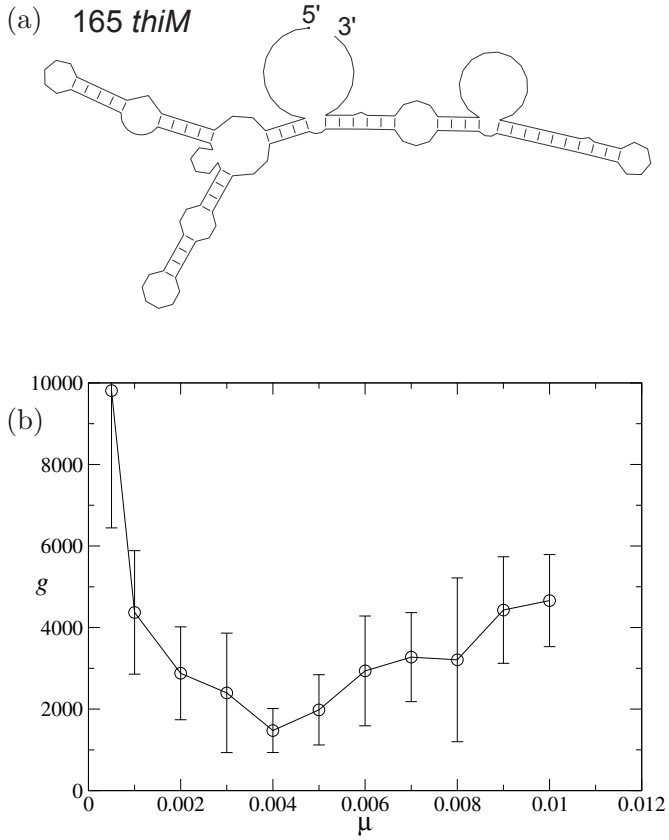


Fig. 2. RNA model. (a) Target structure 165 *thiM*. The molecule starts at 5′ and finishes at 3′. Paired bases are indicated by a short dash. (b) RNA model. Time to find target structure as a function of μ ($N = 1000$, $\beta = 1.0$, averages over 5 realizations were performed). In all figures to be shown, error bars stand for the dispersion σ of a set of $i = 1, \ldots, M$ measurements x_i, with average $\bar{x} = M^{-1} \sum_i x_i$ and $\sigma^2 = M^{-1} \sum_i (x_i - \bar{x})^2$. Averages over time intervals are performed accordingly.

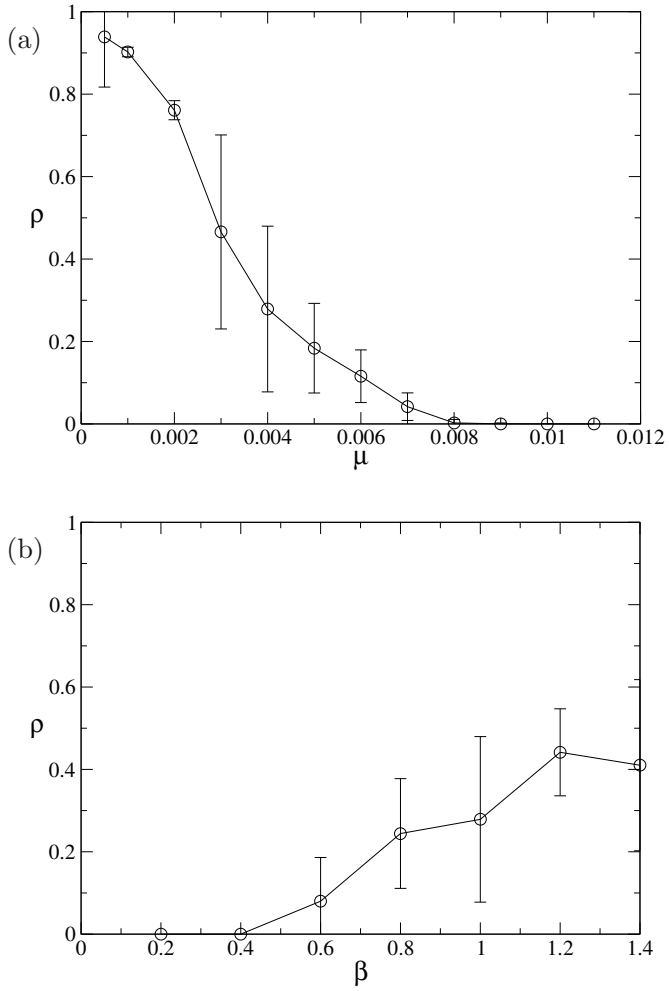


Fig. 3. RNA model. Density of correctly folded sequences as a function of μ (a) and β (b). Parameters: $N = 1000$, $\beta = 1.0$ (a), $\mu = 0.004$ (b). Averages over 5 realizations and 2000 generations in the asymptotic regime were performed.

where d is the average distance of the population to the target structure, $d = \sum_{i=1}^{N} d_i/N$, and $Z = \sum_{i=1}^{N} \exp(-\beta d_i/d)$, where the parameter β denotes the selection pressure.

3.2 Evolutionary population dynamics

Before we describe the phylogenetic properties associated with the evolution of RNA populations, let us present some general features of its evolutionary dynamics. The initial ensemble of random sequences evolves through discrete generations. After a number of generations g (which differs from realization to realization), it finds the target structure for the first time and later reaches an asymptotic state, characterized by statistically stationary quantities. The value of g, and hence the duration of the transient before attaining the asymptotic state, depends on the parameters of the system, especially on the mutation rate μ, as shown in Figure 2b. In accordance with previous work for shorter RNA molecules [29], we observe that the search process for small mutation rates is slow, for intermediate rates fast, and for large mutation rates slightly slower than in the intermediate range.

A relevant quantity to characterize the state of the population, in particular its evolutionary success, is the fraction ρ of sequences folding into the target structure. Due to the stochastic nature of evolution, this quantity fluctuates in time even after reaching the asymptotic regime. Therefore, within this regime, we perform averages over long time intervals (and different realizations, starting from distinct initial RNA populations). In Figure 3a, we display how the average value of ρ varies as a function of μ. For low mutation rates, a large fraction of molecules fold into the target structure. As μ increases, ρ decreases monotonically until it approaches zero. Then, the so-called phenotypic error threshold is crossed, and the mutation rate becomes too large to allow the fixation of the target structure within the population. Also these results agree qualitatively with those obtained for shorter RNA molecules [29]. In Figure 3b, we show the behavior

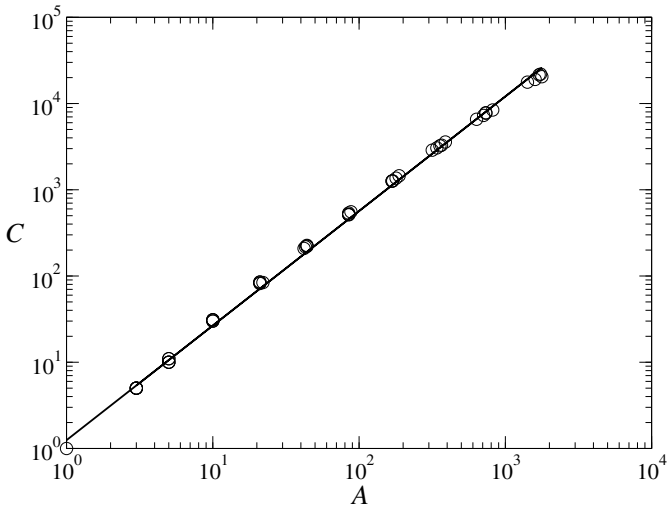


Fig. 4. RNA model. Cumulative branch size C as a function of branch size A. Binned results for 5 trees for the parameters $N = 1000$, $\beta = 1.0$, $\mu = 0.004$ (open circles), together with one regression curve (solid line) are shown.

of ρ as we vary β. Weak selective pressures make evolutionary success difficult (small ρ) while large values of β lead to larger ρ. The parameter β establishes the relative advantage of one variant with respect to the rest of types. For the same mutation rate, increasing the value of β gives a larger relative advantage to structures closer to the target, so the value of ρ increases.

3.3 Phylogenetic properties

Once the population has reached its statistically stationary state, we construct the corresponding phylogenetic tree. Then, we can calculate the subtree branch size A and the cumulative branch size C. For all parameter sets studied, the functional behavior of $C = C(A)$ seems to follow approximately a power law (similarly to what is observed in real data [8]). Further below, we discuss the limits of this approximation. For each simulation, the scaling exponent η is determined by a least-squares regression. An example is displayed in Figure 4. The leaves of a tree contribute with N points $(A, C) = (1, 1)$ to the data, the next branching level with many points $(A, C) = (3, 5)$, almost independently of the overall branching properties of the tree (cf. Fig. 1). Since we are interesting in the scaling properties which become clearer for large values of A and C, we bin the values of A and C in boxes of powers of 2 and in this way avoid a statistical overrepresentation of the low-level branching parts of the trees. In the figure, we show the results of 5 independent realizations (5 trees) and an example of a single power-law fit. The fit values have been averaged, yielding, e.g. for $\mu = 0.004$ a mean exponent of $\eta \approx 1.336$.

In Figure 5a, the dependence of η on μ is displayed. We see that η seems to increase with μ although standard deviations (of the distribution of the 5 average values) are large, in particular for large mutation rates. Still, it must

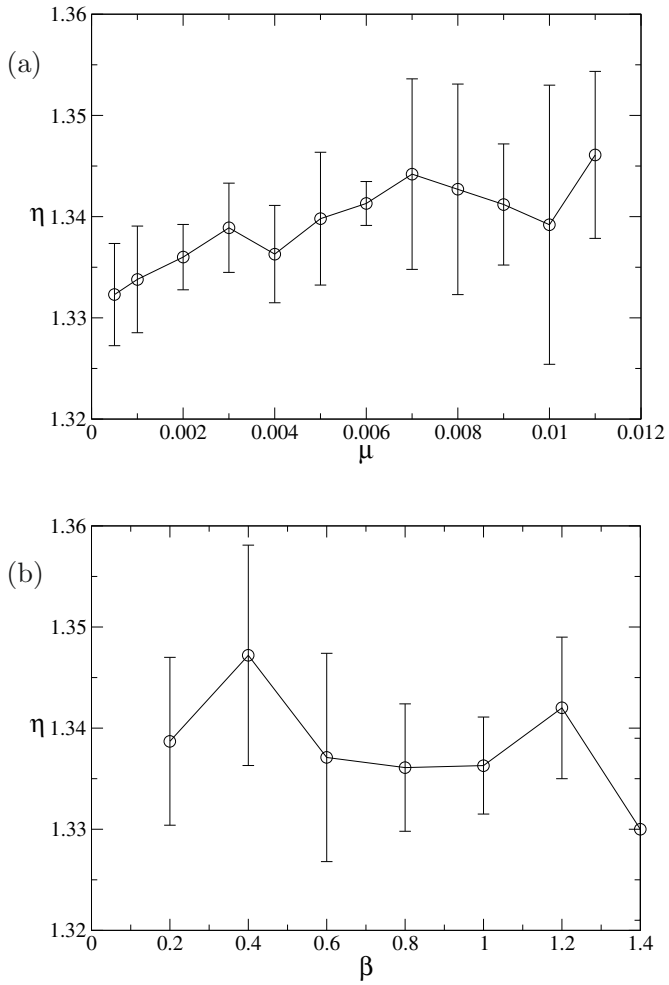


Fig. 5. RNA model. Scaling exponent η as a function of μ (a) and β (b). Parameters: $N = 1000$, $\beta = 1.0$ (a), $\mu = 0.004$ (b). Averages over 5 realizations were performed.

be noticed that the variation of η over the whole range is quite small (mean values from 1.332 to 1.346), taking into account that the evolutionary success, and hence the phenotypic structure of the population, change strongly across the regimes (Fig. 3). If we vary β (Fig. 5b), we observe a less clear trend, although also here the mean values of η remain within a relatively narrow range.

4 Simple model with inheritance and mutation

The RNA evolutionary model studied in the previous section, albeit biologically grounded and simple from a biological point of view, is already complex from a theoretical or computational point of view due to sequence-structure map. In order to understand to which extent the results obtained are generic, we introduce a phenomenological model defined by simple rules but still containing basic evolutionary mechanisms.

4.1 Evolutionary algorithm

Consider a set of $i = 1, \ldots, N$ individuals, each characterized by a fitness $f_i(g)$ at generation g. At generation $g+1$, a group of N new individuals substitutes the previous one. The probability $p(i \to j)$ that individual j at generation $g+1$ originated from individual i at generation g is proportional to $f_i(g)$,

$$p(i \to j) = \frac{f_i(g)}{\sum_k f_k(g)}, \tag{12}$$

where the sum runs over the whole population. With probability ν, the fitness of the offspring takes a value randomly drawn between 0 and 1, i.e. $f_j(g+1) = \delta$, $\delta \in]0,1[$; with the complementary probability $1-\nu$, it inherits the fitness of the parent i, $f_j(g+1) = f_i(g)$. In the limit, $\nu \to 1$ the system becomes memoryless and there is no correlation between the reproductive rates of the ancestors and the rates of the current descendants. In the limit $\nu \to 0$ the population is completely correlated. However, after a transient period the initial diversity of fitness values is lost in the latter case, since only one of the initial individuals becomes the ancestor of all of the extant group. At that point, there is no selection and the model becomes effectively neutral, thus equivalent to ERM. We will refer to this model as IM model (for inheritance with mutation).

The complexity of the RNA evolutionary model studied in the previous section requires long computational times. Thus, both the population size N and the number of realizations for each value of the parameters were limited to relatively small values. The IM model permits to work with larger systems (we will show results for trees with up to 10^4 tips) and to perform averages over a larger number of independent realizations, thus obtaining better estimates of relevant quantities.

4.2 Evolutionary population dynamics

This model differs from the RNA model in that there is no target function to drive the evolution of the population (a condition analogous to evolving towards a target RNA structure), and in consequence there is no error threshold. From a biological point of view, the interesting duality of having a meaningful representation of both genotype and phenotype, reflected by RNA sequence and secondary structure, is lost in this simple model.

A way to quantify the degree of optimization of the population is to calculate the average fitness $\langle f \rangle$ in the asymptotic regime, averaging over all individuals, long time intervals and different realizations. Figure 6 illustrates that if the mutation rate is low, selection drives the population into a state with many individuals with high fitness, while for a high mutation rate, fitness values are essentially random and hence take an average value 0.5. In terms of fitness landscapes, we are implementing a single-peak landscape where the population clusters close to the fitness maximum for low μ, while it spreads steadily as μ increases. In spite of the fact that the IM model is much simpler, the average fitness shows a behavior qualitatively similar to the density ρ of correctly folded sequences in the RNA model (cf. Fig. 3a).

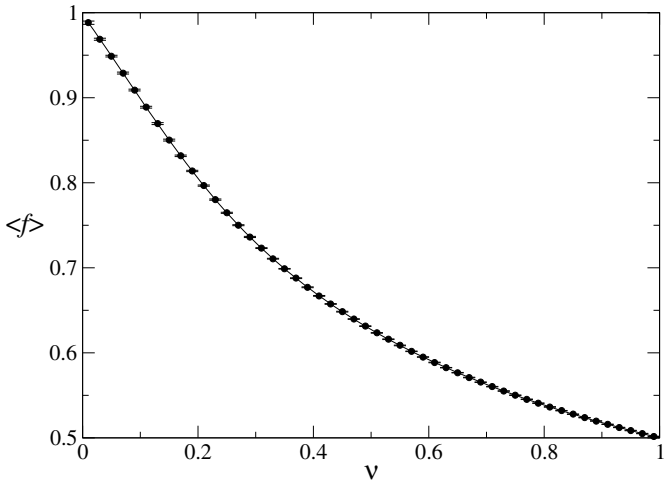


Fig. 6. IM model. Average value of fitness as a function of the mutation rate ν. The system size is $N = 1000$, and averages over 100 independent trees for each value of ν were performed.

4.3 Phylogenetic properties

After reaching the asymptotic regime, we can build the phylogenetic tree, calculate A and C, and determine η. In Figure 7a, we show η as function of ν and observe that in spite of the completely different evolutionary outcomes of the cases $\nu \to 0$ and $\nu \to 1$, the scaling exponent η does not vary much.

Next, we have studied how η depends on the system size N. Before we actually show the results for the IM model, let us first discuss what we observe for the backward coalescent model (as introduced in Sect. 2). There, trees are constructed backwards by joining two species at each generation. As we have discussed in the Introduction, and shown in Section 2, the functions $P(C)$ and $C(A)$ in the completely balanced tree have logarithmic corrections to their scaling behavior that especially affect small trees. In our analysis of the dependence of the exponent η on the system size, we observe for the IM model a decrease in η as trees become larger, as represented in Figure 7b. Furthermore, we detect a clear correspondence between the behavior observed in the evolutionary IM model (dots) and that of a completely uncorrelated phylogeny, represented by the coalescent model (solid line). Compared to the coalescent, the IM model maintains certain degree of correlations that yields values of η above those of the coalescent model for the system sizes explored. Nevertheless, its behavior closely follows that of the ERM class for which we know that in the limit $N \to \infty$, the exponent $\eta \to 1$.

5 Comparison of the scaling properties

In this section we review and quantitatively compare the scaling properties of the phylogenies generated by four of

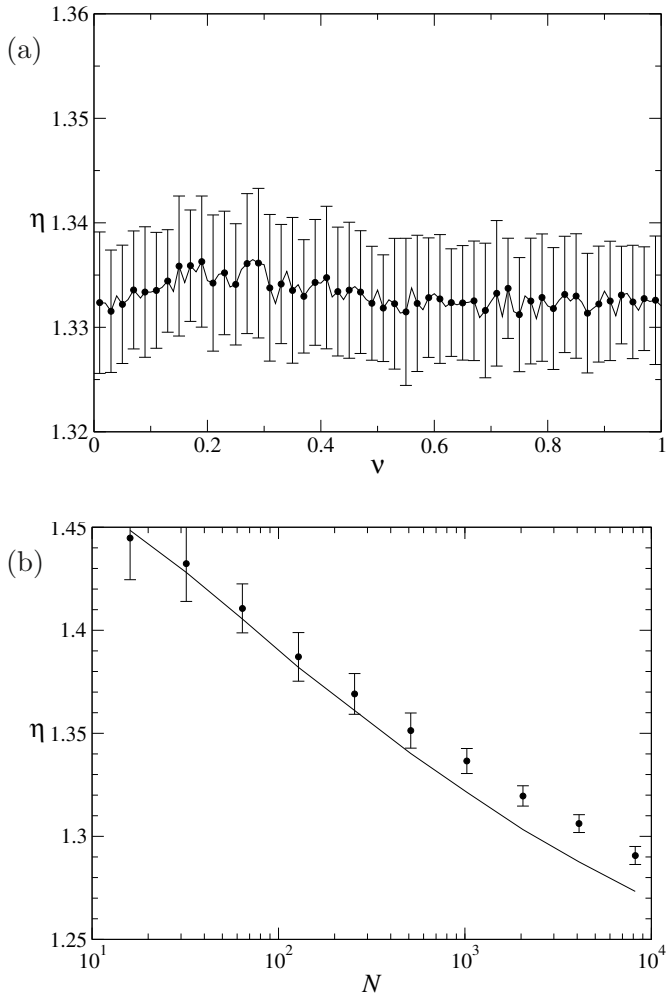


Fig. 7. IM model. (a) Average value of the scaling exponent η as a function of ν. The system size is $N = 1000$, and averages over 100 independent trees for each value of ν were performed. (b) Average value of the scaling exponent η (dots). The mutation rate is $\nu = 0.2$, and averages over 100 independent trees for each value of N were performed. The solid line stands for the exponent η obtained from averages over 10^3 independent trees of corresponding size generated with a backward coalescent model.

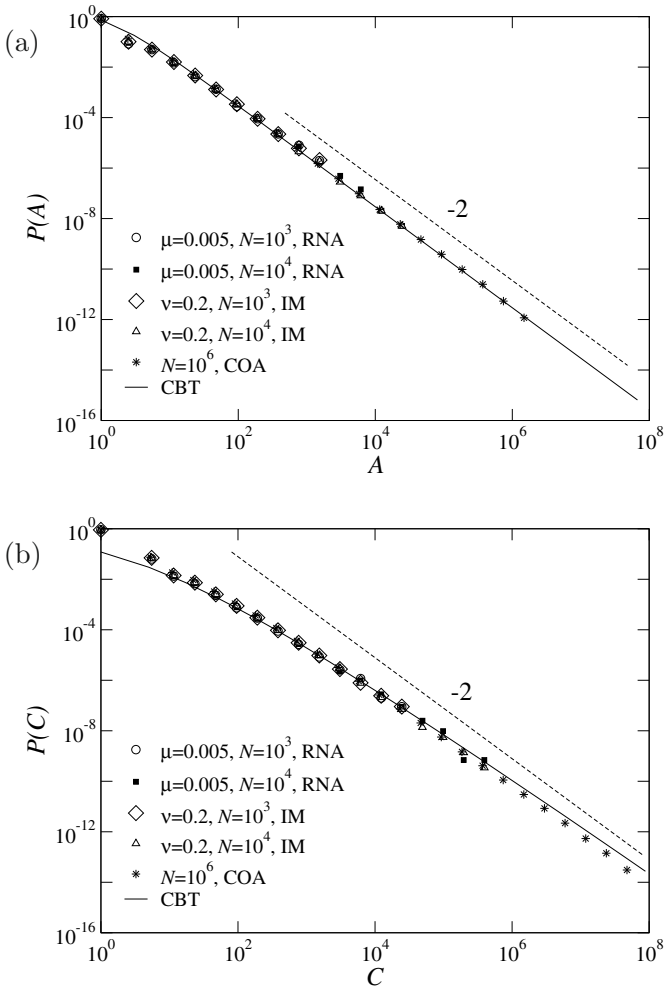


Fig. 8. Distributions of topological quantities for the four evolutionary models discussed in this work. (a) Subtree branch size distribution $P(A)$. (b) Cumulative branch size distribution $P(C)$. The legend specifies which symbols correspond to each of the models and the parameters used. In the case of RNA, the selection pressure $\beta = 1$. Averages over 5 (RNA), 100 (IM), and 50 realizations (COA) have been performed. Dashed lines have the slopes -2, corresponding to the scaling exponents α and γ for a completely balanced tree in the asymptotic regime.

the evolutionary models discussed in this work. In decreasing degree of biological complexity (thus realism) they are (i) evolution of RNA populations with explicit selection on the phenotype and two evolutionary parameters, mutation rate μ and selection pressure β, analyzed in Section 3 (RNA model); (ii) simple model with individuals characterized by their fitness and one evolutionary parameter, the mutation rate ν, presented and studied in Section 4 (IM model); (iii) the coalescent, implemented as a set of extant species whose phylogeny is reconstructed backwards in time by randomly selecting two of the remaining species at each generation and merging them in a common ancestor (COA model); and (iv) a fully symmetric and balanced tree, whose scaling properties are known and have been rederived in Section 2 (CBT model).

In the next two figures we compile the results obtained for the topological quantities $P(A)$, $P(C)$ and $C(A)$, as defined in Section 2, corresponding to the four models enumerated. They summarize our results and support our main conclusion: simple evolutionary processes generate phylogenetic trees with topological properties essentially indistinguishable from the ERM class. Figure 8a presents the distribution $P(A)$ for the four models above and two different system sizes for RNA and IM. We recall from Section 2 that, for CBT, $P(A) \propto A^{-2}$. Thus, as the size of the trees increases, the slope of the distribution for CBT (solid line) approaches -2, as indicated by the dashed line in the plot. Since the COA model belongs to the ERM class, the corresponding curve also has an asymptotic slope of -2. More strikingly, also the evolutionary models (RNA and IM) seem to follow this behavior. Thus, though the

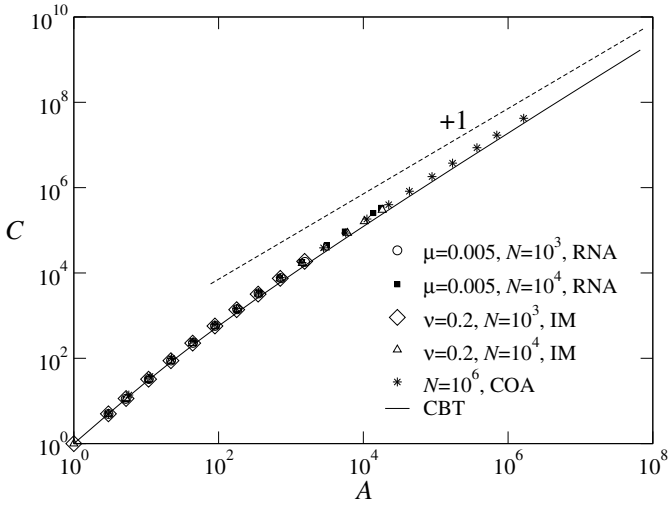


Fig. 9. Cumulative branch size C as a function of the branch size A for the four evolutionary models discussed in this work. Parameters as in Figure 8. The dashed line has slope one, corresponding to the scaling exponent η for a completely balanced tree in the asymptotic regime.

analytic results could only be derived for the fully symmetric case of CBT, our results support the view that the dominant functional forms of the phylogenetic trees obtained through the different models here studied asymptotically agree with the CBT.

The corresponding distributions $P(C)$ are compared in Figure 8b. There is a clear change in the scaling in this case, since a visible bending affects the whole range of C-values explored. At odds with other systems (as transportation networks [19]), where the functional form of the distribution of accumulated branch sizes seems to be dominated by a pure power law, we have shown for the class of CBT that $P(C) \propto (C^2 \ln C)^{-1}$. Furthermore, the logarithmic correction seems to be shared by the phylogenetic trees arising from all the models analyzed. As is clearly seen in the figure, attemps to fit the distribution $P(C)$ with a pure power law may yield misleading results.

The complex scaling behavior of $P(C)$ is also reflected in the relationship between C and A, as shown in Figure 9. The analytical results for CBT show that $C \propto A \ln A$, again with a logarithmic term that causes a systematic deviation from a pure exponent $\eta = 1$ in all the range of tree sizes that could be explored. Also here we observe that the evolutionary models over a large range of tree sizes qualitatively agree with the results obtained for completely balanced trees.

6 Discussion

Intrinsic evolutionary parameters and environmental conditions determine the fate of species, their ability to survive and radiate, and the eventual size of their clades. In this work we have addressed the question how these processes modify the topology of phylogenetic trees. We have used models of individual replicators evolving towards an optimal target function (RNA model) or according to a fitness function (IM model), to create phylogenetic trees and subsequently investigate the scaling properties of their topological quantities. We have shown that, in evolutionary models with different degrees of complexity, finite-size effects result in quantitative changes in tree topology that largely exceed those due to mutation and selection.

It is indeed remarkable that parameters as the selection pressure and the mutation rate play such a weak role in the topological properties of phylogeny, while the size of trees significantly affects the measured values of the scaling exponents. If, as hypothesized (see, e.g. Ref. [16]), mutation acts in the sense of erasing correlations as time (i.e. tree size) increases, it can be expected that smaller systems are more correlated, hence present a higher degree of imbalance and as a consequence yield larger values of η: they are by construction closer to imbalanced trees, for which in the limit of complete imbalance $\eta = 2$. Larger values of η for smaller trees is actually what we observe, together with an important decrease of η for increasingly large systems, a variation much larger than that due to changes in the mutation rate.

In this work, we have considered large trees and focused on the scaling behavior of the subtree size A and cumulative branch size C. The effects presented here do not contradict findings for small trees where tree imbalance is generic: evolutionary trees produced by a Moran's model are found to be only slightly more imbalanced than neutral ones [23]. We emphasize again that the bending of the distributions demonstrates that imbalance of small trees is compatible with the ERM scaling.

We should mention that we have not studied the scaling behavior of trees with persistent imbalances. There are some models in the literature where branching probabilities of species are assigned according to their position in the tree (cf. on the size of the parent clade), thus causing persistent asymmetries or imbalances [14]. That class of models has not yet been analyzed from the viewpoint of the asymptotic scaling of C and A. This might be an interesting objective for future investigations.

A way of distinguishing whether the non-trivial exponents measured in natural phylogenies genuinely reflect non-trivial aspects of the evolutionary process itself or, on the contrary, result from the small size of the trees considered, would be to check for the presence of correlations between the measured values of α, γ, and η and the number of species in each tree. Even in the case that those correlations would be weak or absent in real systems, we believe that other quantities beyond the topological properties studied here are necessary to characterize the role of different mechanisms shaping the tempo and phylogenetic structure of the evolutionary process. In the light of our results, we can but agree with previous investigations leading to the conclusion that the presence of universal scaling exponents can be considered just a consequence of the parent-child structure of a taxonomy [32,33].

The authors acknowledge conversations with E.A. Herrada, E. Hernández-García and V.M. Eguíluz who draw our attention

to the problem tackled in this work. Financial support from Spanish MICIIN through project FIS2008-05273 is gratefully acknowledged.

References

1. P.H. Raven, B. Berlin, D.E. Breedlove, Science **174**, 1210 (1971)
2. C. Darwin, *On the origin of species by means of natural selection* (John Murray, London, 1859)
3. J.C. Willis, G.U. Yule, Nature **109**, 177 (1922)
4. G.U. Yule, Phil. Trans. Roy. Soc. London B **213**, 21 (1925)
5. D. Aldous, Stat. Sci. **16**, 23 (2001)
6. B. Burlando, J. Theor. Biol. **146**, 99 (1990)
7. D. Aldous, M. Krikun, L. Popovic, J. Math. Biol. **56**, 525 (2008)
8. E.A. Herrada, C.J. Tessone, K. Klemm, V.M. Eguíluz, E. Hernández-García, C.M. Duarte, PLoS ONE **3**, e2757 (2008)
9. F. Delsuc, H. Brinkmann, H. Philippe, Nat. Rev. Gen. **6**, 361 (2005)
10. F.D. Ciccarelli, T. Doerks, C. von Mering, C.J. Creevey, B. Snel, P. Bork, Science **311**, 1283 (2006)
11. A.O. Mooers, S.B. Heard, Quart. Rev. Biol. **72**, 31 (1997)
12. *TreeBASE*, `http://www.treebase.org`
13. M. Kirkpatrick, M. Slatkin, Evolution **47**, 1171 (1993)
14. M.G.B. Blum, O. François, Syst. Biol. **55**, 685 (2006)
15. V. Savolainen, S.B. Heard, M.P. Powell, T.J. Davies, A.O. Mooers, Syst. Biol. **51**, 835 (2002)
16. B. Derrida, L. Peliti, Bull. Math. Biol. **53**, 355 (1991)
17. J. Aguirre, S.C. Manrubia, Phys. Rev. Lett. **100**, 38106 (2008)
18. W. Hordijk, J.F. Fontanari, P.F. Stadler, J. Phys. A **36**, 3671 (2003)
19. J.R. Banavar, A. Maritan, A. Rinaldo, Nature **399**, 130 (1999)
20. D. Garlaschelli, G. Caldarelli, L. Pietronero, Nature **423**, 165 (2003)
21. J. Camacho, A. Arenas, Nature **435**, E3 (2005)
22. P. Donnelly, S. Tavaré, Annu. Rev. Genet. **29**, 401 (1995)
23. L.P. Maia, A. Colato, J.F. Fontanari, J. Theor. Biol. **226**, 315 (2004)
24. T.E. Harris, *The Theory of Branching Processes* (Springer-Verlag, 1963)
25. P. De los Rios, Europhys. Lett. **56**, 898 (2001)
26. R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, D.E. Knuth, Adv. Comp. Math. **5**, 329 (1996)
27. M. Eigen, Naturwissenschaften **58**, 465 (1971)
28. P. Schuster, P.F. Stadler, Comp. & Chem. **18**, 295 (1994)
29. M. Stich, C. Briones, S.C. Manrubia, BMC Evol. Biol. **7**, 110 (2007)
30. I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, P. Schuster, Monatsh. Chem. **125**, 167 (1994)
31. W. Winkler, A. Hahvi, R.R. Breaker, Nature **419**, 952 (2002)
32. G. Caldarelli, C.C. Cartozo, P. De los Rios, V.D.P. Servedio, Phys. Rev. E **69**, 035101(R) (2004)
33. A. Capocci, F. Rao, G. Caldarelli, Europhys. Lett. **81**, 28006 (2008)