UNIVERSITAT DE LES
ILLES BALEARS

Master Thesis

Simple Branching Models for Macroevolution

Author:
Murat Tuğrul

Advisor:
Emilio Hernández-García
Co-Advisor:
Víctor M. Eguíluz

IFISC
Instituto de Física Interdisciplinar y Sistemas Complejos

August 28, 2009

Simple Branching Models for Macroevolution

by

Murat Tuğrul

THESIS

Presented to

The University of Balearic Islands

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN PHYSICS

September 2009

The scientist does not study nature because it is useful; he studies it because he delights in it, and he delights in it because it is beautiful. If nature were not beautiful, it wouldn't be worth knowing, and if nature were not worth knowing, life wouldn't be worth living.

Henry Poincaré

to Life..

Abstract

The theoretical tools of Statistical Physics offer powerful techniques to analyse problems in which chance and probabilities play a role. One of such subjects is the understanding of the rules of macroevolution, i.e., evolutionary development and diversification of species, which remains a not well developed part of evolutionary biology.

Phylogenetic trees, describing the estimated evolutionary relationships between biological species, are obtained directly from molecular data and are an important indirect evidence for diversification patterns in macroevolution. Therefore, analysing the structures of such estimated trees and comparing to those obtained from branching models is an interesting approach to capture the rules of macroevolution.

In this thesis, we analyze the phylogenetic trees in the TREEBASE and PANDIT databases and characterize their topology (in particular their balance degree) via the *mean depth*, i.e. the average number of ancestor nodes from the tips to the root. A non-logarithmic scaling with tree size is found, which is not easy to get with branching models existing in the literature. With this motivation, we analyze analytically and numerically three simple branching models, two of which are proposed by us, and try to find their biological meaning if possible. The first is *Ford's alpha model*; although a power law scaling of the mean depth with tree size was established analytically, our numerical results illustrate that the asymptotic regime is approached only at very large tree sizes. For the second model, named as *activity model*, we show analytically and numerically that it also displays a power law scaling of the mean depth with tree size at a critical parameter. Finally, we propose the so called *age model* in which the probability of branching depends on the age of the tips with a power parameter. The results of this model at a critical parameter value, which we are capable to express analytically, display a scaling behavior similar to the one obtained from databases analysed. In addition it is potentially open to biological interpretation.

Contents

Part I

Introduction

1

Background

1.1 Biological Evolution

After centuries of human performance on the domestication of animals and the cultivation of plants, and after many decades of very detailed scientific observations and documentations of the shapes and the behaviors of living organisms, it was started to be discussed that the forms of living matter do not persist unchanged but, on the contrary, they are in evolutionary process at time scales much longer than individuals' lifetime. The theorizing work of this biological evolution took place around 150 years ago and was due to Darwin and Wallace by asserting that the evolution is governed by the *variations* of the characters and their relative *selection* (or equivalently *election*) by a complex dynamical system of both living and nonliving matters, namely by *Nature*, due to any advantage they gained over the others (Darwin, 1859).

This theory of evolution, especially during the first decades of the last century, was combined with genetics shaping the classical paradigm of biological evolution until present (Huxley, 1942). In recent years, mainly due to rapid increase of the genetical data and their interpretations, the new theoretical and experimental perspectives in science relating to biological evolution have begun to emerge. To give some recent examples, Roughgarden et al. (2006) have brought a new argument for the sexual selection which is opposing to Darwin's sexual selection. Some groups like Goldenfeld and Woese (2007), Vetsigian et al. (2006), Frigaard et al. (2006) have studied, from new points of view, the lateral gene transfers and their impact on evolution (Syvanen, 1985). Balcan et al. (2007) proposed an informational theoretic model, which mimics with high accuracy the global structural properties of the network of transcriptional regulatory interactions found in yeast organism, which can infer the informational (or entropic) aspects of the biological evolution (Volkenstein, 1994) in the gene coding.

In short, new ideas and studies might be apt to improve our knowledge about biological evolution and even may change today's paradigm in the near future, possibly with the approaches from other disciplines, but especially from physics.

1.2 Physics & Evolution

Until recently improvements in physics and biology have occurred quite independently. Therefore, dissimilar *modus operandi* have been developed which hardened the communication between two disciplines. However, today with increasing interest in biological knowledge, it is expected that such dissimilarity decreases and the experience of physics contributes to gain insight into the biological problems. Here, we try to explain how these contributions from physics to biological evolution can be and give some examples from the literature.

First of all, in some cases, the forms of the living organisms might be dictated only by the nonliving part of nature. For example, Barrio et al. (1997) and Aragon et al. (2002) have studied the organisms whose form is pentagonally (five-fold) symmetric, such as sea urchin. They brought a physical explanation to this five-foldness by investigating the nonlinear Turing equations, which are studied in pattern formation. Another example, that fits here well though very recent, is regarding the distinct number of nucleic acids in DNA and of amino acids in proteins, which are 4 and 20, respectively. Note that each amino acid is coded by codon (the code consists of three nucleic acids). Common biologist's explanation to these numbers is by means of *frozen accident* during the earlier epoch of the evolution. However, Patel (2003) argues that Grover's algorithm discussed in quantum physics gives the optimum solution which coincides in 1 and 3 quantum queries (Pusuluk, 2009). See also Ref. (Lloyd, 2009) for a review about other studies relating Quantum Physics and Evolution.

Secondly, an analogy between the evolutionary system and a well understood physical system can be considered. If enough completeness is obtained at the analogy, the tools developed in physical system can be applied to evolutionary system. For example, Sella and Hirsh (2005) discuss the formal resemblance of the population genetics and statistical physics. In Table 1.1, we give their analogy between evolutionary dynamics and statistical physics. In the same line, one can also see Refs (Dietz, 2005; Kowald and Demetrius, 2004) for an attempt to construct *the evolutionary entropy.*

Lastly and also as the main philosophy behind this thesis research, the usual approach of physics to the problems, which is the construction of complexity from basic proposed units/interactions and a thorough investigation/comparison at each constructed level, can be applied to evolutionary system. In other words, at first the lower level laws of evolutionary dynamics might be proposed despite the complexity of system. Later, the system is brought to higher levels with the help of mathematics and/or computers. At the end, the system is subjected to comparison with evidences from the real system. For example, Pigolotti et al. (2007) have studied some patterns of evolutionary ecological system by assuming that the populations are localized in an abstract niche space and using a Lotka-Voltera type interaction dynamics. Similar examples for building the complexity in the context of evolution by physicists can be found in Refs (Higgs and Derrida, 1992; Powell and McKane, 2008).

Table 1.1. The analogy between evolutionary dynamics and statistical physics given in Ref. (Sella and Hirsh, 2005). E_i for energy of the state i, N for population size; k_B for Boltzmann constant; T for temperature; ν for analogous term for β; P^i probability of the state i; G for free energy; f_i for fitness; x for additive fitness.

Object	Evolutionary Dynamics	Statistical Physics
State variable	$\vec{g} = (A, C, T..G, A)$	$\vec{s} = \{(\vec{q_k}, \vec{p_k})\}$
Additive fitness and energy	$x = ln(f(\vec{g}))$	$E = \hat{H}(\vec{s})$
Population size and temperature	$\nu_{Moran} = N - 1$ $\nu^h_{WF} = 2(N - 1)$ $\nu^d_{WF} = 2N - 1$	$\beta = 1/k_B T$
Boltzmann factor	$P^i_{s.s.} \sim e^{-\nu(-x_i)}$	$P^i_{eq} \sim e^{-\beta E_i}$
Invariance	$f_i \rightarrow C f_i$	$E_i \rightarrow E_i + C$
Free fitness and free energy	$G = \langle x \rangle + \frac{1}{\nu} H$	$-G = -(\langle E \rangle - \frac{1}{\beta} H)$
Equilibrium scale	$\nu(x_j - x_i) = 1$	$\beta(E_j - E_i) = 1$

1.3 Macro- versus Micro-Evolution

Since the evolutionary process of living beings can be thought in different time/space scales and organization levels, one can speak about the evolution of biomolecules such as proteins, of cells, of organs such as eyes, of populations, of species, of genera, etc., therefore, the field of biological evolution can be separated into many subfields. However, we see that traditionally the research field of biological evolution has been divided into two main subfields in the literature: micro- and macro-evolution (See Ref. (Erwin, 2000) for details of the history).

Microevolution, at one hand, deals with the changes occurring within a species or population in comparably short time and often refers to population genetics or evolutionary genetics. It should be noted that many extensive and systematic microevolutionary research have been done mainly due to the availability of empirical studies and some mathematical improvements have been already supplied such as the neutral theory of molecular evolution (Kimura, 1983).

On the other hand, macroevolution is the long-term development and dynamics of the formation (speciation) or the extinction of species and higher classification units. Not only due to the problems of a universal definition of species (Hendry, 2009) but also due to the difficulty (generally speaking: impossibility) of doing experiments, this part of evolutionary studies did not improve as much. It had depended on the fossil records which are at best limited to a small portion of the living classifications. Accordingly, accepted mathematical theories are still missing in literature.

The recent technological improvements in the last decades have formed a new branch of biology: *Molecular Phylogeny.* It supplies us evidences of macroevolutionary relationships of species obtained from molecular analysis. It is expected to that these phylogenetic studies result in a better elucidation of the macroevolution.

1.4 Thinking Macroevolution as Tree-like and Phylogenetic Trees

The real history of the speciation on Earth was considered by Darwin (1859) (see Figure 1.1), or even in earlier times by Lamarck (Gregory, 2008), to be branching tree-like where the tips are extant species and the internal nodes are the ancestor species. Today, this hypothesis of tree-like macroevolution is still highly used and we follow this consideration in our research. However, the reader should be warned that there exist phenomena such as the hybridization of species (Hendry, 2009) and lateral gene transfer (Syvanen, 1985) which make the true representation of the macroevolution like a network[1] rather than a tree.

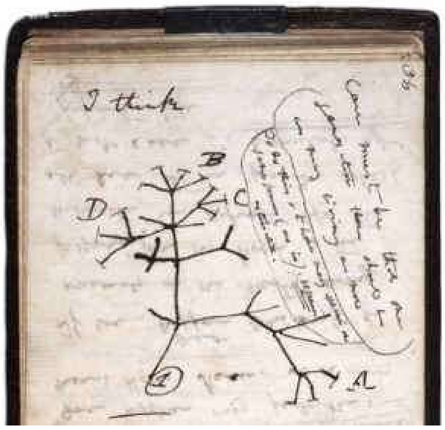

Figure 1.1. An evolutionary tree sketched by Darwin ($\sim$ 1837) found in one of his note books. It might be noted that the unique illustration included in *On the Origins of Order* was again an evolutionary tree.

Therefore, we seek this tree-like structure of the evolutionary relationships of species. As mentioned above the phylogeny reconstructs trees of the evolutionary relationship of a number of existing species, named as phylogenetic trees, by using their molecular data, i.e., DNA sequences. There are different types of phylogenetic trees, two of which are *cladograms* and *phylograms.* A cladogram is simply a representation of the order of the evolutionary closeness of different taxa (a taxonomic group of any rank, such as a species, family, or class.). Let me explain some important information it carries by using the example cladogram shown in Figure 1.2. The taxa **A** and **B** are more closely related in the set of taxa **{A, B, C}**. Left and right

[1] Generally, the word *web* is used in the evolutionary biology literature.

edge ordering does not make a difference (Aldous, 1996). Tips are extant species whereas internal nodes are inferred speciation events. For the sake of simplicity, one can read them as the ancestor species. Therefore, the *most common ancestor* of **A** and **B** is the internal node with one asterisk, whereas the one with two asterisk is of **A**, **B** and **C**; **F** is the most distantly related one and named as *outgroup*. Most of the time, the outgroup is necessary to root the tree (Gregory, 2008). Actually, there are also unrooted evolutionary trees but in this research we are not taking them into account and the databases we use include only rooted trees.

Phylograms differ from cladograms by having branching lengths, usually indicating the real time or the amount of change in the sequences. In this research, we only use the cladogram representation although some of our studies can be related to phylograms and can be studied further in this context.

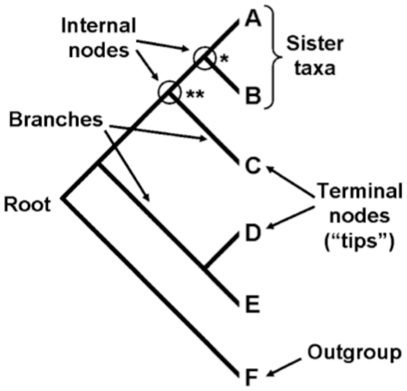

Figure 1.2: Basic terms of phylogenetic trees are displayed (Gregory, 2008).

There are many databases of reconstructed (estimated) evolutionary cladogram trees available today. In this research, we use the TREEBASE[2] and PANDIT[3] databases which are public repositories. Figure 1.3 shows one example from TREEBASE. It should be emphasized that due to resolution in reconstruction techniques there might appear polytomies (spliting to more than 2 branches) in estimated trees, however, in our modeling we concentrate on only binary trees as other studies in the literature.

It is important to note that there are discussions of the possible non-biological effects to reconstruction results such as artifacts of estimation techniques, the incompleteness of the trees, the decisions of phylogeneticists, etc. (See References (Barraclough and Nee, 2001; Mooers and Heard, 1997) for a detailed discussion). However, we leave these discussions out of our research and see the phylogenetic trees as an important indirect evidence of diversification structure throughout the evolutionary history on Earth. It is believed that scrutinizing the structure of phylogenetic trees might be very useful to elucidate the nature of macroevolution (Barraclough and

[2] http://www.treebase.org/

[3] http://www.ebi.ac.uk/goldman-srv/pandit/

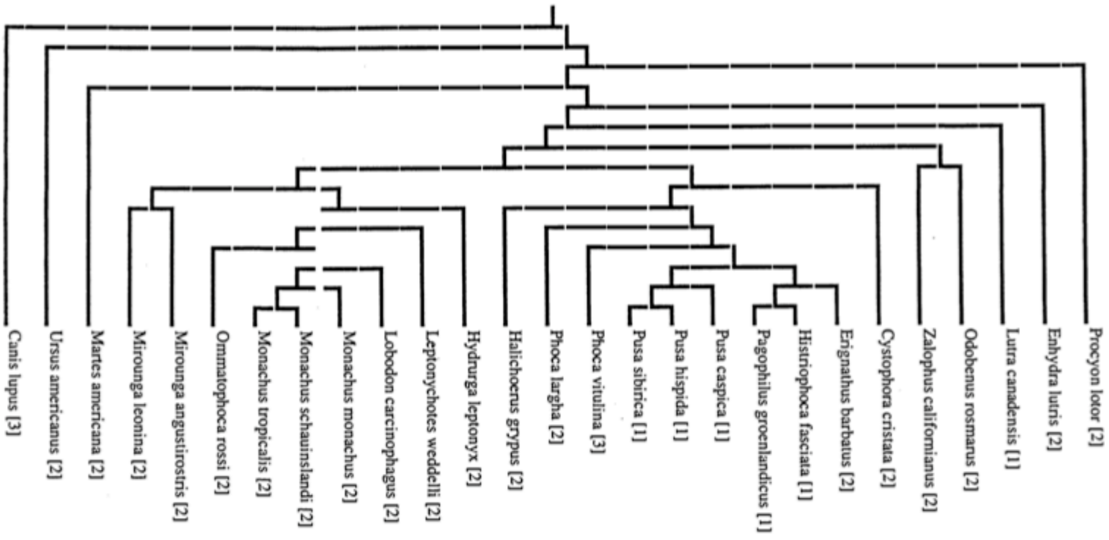

Figure 1.3: An example tree (Aldous, 2001a) from TREEBASE.

Nee, 2001; Reznick and Ricklefs, 2009; Ricklefs, 2007). This brings us to discuss the necessary tools to study the structure of the trees.

1.5 Tree Balance: *Mean Depth*

From the mathematical point of view, a tree is a graph (network) without cycles. By this property many of the topological measurements available from graph theory, e.g. clustering and degree distributions, give no information here. One of the notions which we can relate to the topology of the tree is its balance, i.e., left-right symmetry for branching. For a graphical explanation of the balance, see Figure 1.4 for the most extreme two cases in binary trees: most balanced (namely symmetric or cayley) and the most unbalanced (namely comb) trees.

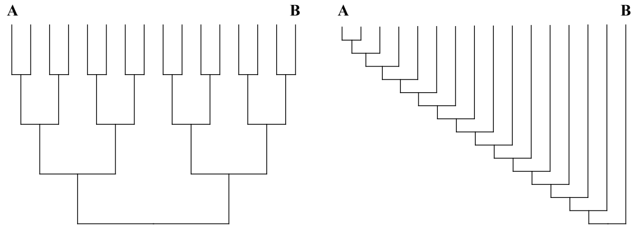

Figure 1.4: The most balanced (at left) and imbalanced trees (at right) with 16 tips.

From the biological point of view, the balance shape might give us some hints regarding how the diversity of life propagates through evolutionary time. In other words, the question: *Does speciation tend to be more in one part of hypothetical tree of life than in other parts?* can be answered by investigating the balance structure.

Several indices have been proposed and used in the literature for a quantitive way to present the imbalance of a tree. References (Agapow and Purvis, 2002; Matsen, 2006; Mooers and Heard, 1997) discuss the comparison of them in detail. However, the main problem of finding a homogeneously distributed index (See Figure 1.5) remains still problematic and to our knowledge, there is no generally accepted one yet (Matsen, 2006).

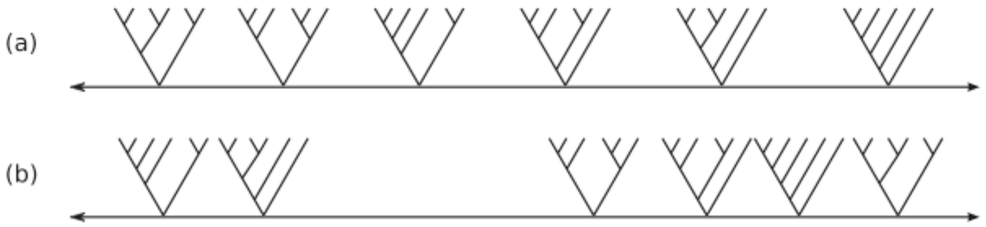

Figure 1.5. A good index of measuring the balance is expected, at least intuitively, more orderly and homogenously distributed on the index line. In this sense index **a** can be said to better than **b** (Matsen, 2006)

In this research, considering 1) the reported comparison of the indices by Matsen (2006) and Agapow and Purvis (2002), 2) the possibility of using them also in polytomies, 3) possibility of analytical derivations from models and 4) the most importantly, for the sake of a clear biological meaning, we use the *mean depth* of a tree: $\langle d \rangle_n$ (Sackin, 1972) being *the average number of ancestor nodes from the tips to the root*, i.e.,

$$\langle d \rangle_n = (\sum_{i=1}^{n} d_i)/n \tag{1.1}$$

where d_i is the depth of the tip i, i.e., its number of ancestors till to the root (inclusive), and n is the total number of tips in the tree.

It should be stressed that the mean depth can not be a direct comparison of balance among any group of trees since it is a quantity related to the shape of tree in longitudinal direction whilst balance is regarding the shape at crossing direction. However, if a set of given size binary (or more in general, k-ary) trees are considered then one can use it for balance measurement. Indeed, we are interested to know how the mean depth of the tree scales with system size n (which is the number of tips). We use the behavior of the scaling to compare the models (producing binary tree) and estimated trees in our research.

Thus, it is important to note here that the mean depth scaling of TREEBASE trees is reported as $\langle d \rangle \sim \log^2 n$ in Ref. (Blum and François, 2006) in context to its comparison with AB model (will be discussed in the following section) and our results will not falsify this peculiar scaling as can be seen in the Original Research part.

Now let us see the exact scaling of the extreme cases for the binary trees, i.e., the most imbalanced and balanced trees, which we need for referencing to trees at the investigation.

1.5.1 Mean Depth Scaling of the Most Imbalanced (Comb) Tree

Note that a binary tree of size n is a comb tree if and only the set of the tip's depth is $1, 2, 3, ...(n-3), (n-2), (n-1), (n-1)$. Therefore,

$$n\langle d\rangle_n = 1 + 2 + 3 + ... + (n-2) + (n-1) + (n-1) \tag{1.2}$$

so dividing each side by n, the exact mean depth is obtained

$$\langle d\rangle_n = \frac{n}{2} + \frac{1}{2} - \frac{1}{n} \tag{1.3}$$

from which the scaling with n is

$$\langle d\rangle_n \sim n \tag{1.4}$$

1.5.2 Mean Depth Scaling of the Most Balanced (Symmetric) Tree

Note that a binary tree is the symmetric (cayley) tree if and only if its size is $n = 2^i$ where $i = 1, 2, 3...$ is the label of the tree levels, and each depth from tips are the same. Therefore, realizing that the mean depth of each level is precisely $\langle d\rangle_n = i$ and using the definition,

$$\langle d\rangle_n = \log_2 n = \frac{\log n}{\log 2} = (1.44...) \log n \tag{1.5}$$

which is said to be scaling with n as

$$\langle d\rangle_n \sim \log n \tag{1.6}$$

1.6 Macroevolution Modeling with Branching Processes

Surprisingly, the modeling of macroevolution was slow to appear in the literature and generally was about the maintenance of genetic variation rather than speciation events (Gavrilets, 2003). The reader might see Refs (Gavrilets, 2003; Powell and McKane, 2008) for some modern examples of modeling macroevolution in contexts other than branching processes discussed here.

Branching processes out of the evolutionary context were also studied extensively and some mathematical foundations were attained in earlier time (Harris, 1963). For examples of non-macroevolutionary branching processes, one can see the following references; (Stevens, 1974) for real trees; (West et al., 1997) for blood vessels; (Rodriguez-Iturbe and Rinaldo, 1997) for river networks; (Klemm et al., 2005, 2006) for computer file systems; (Jain and Dubes, 1988) for classification algorithms; (Kingman, 1982, 2000) for the coalescent theory in population genetics; (Harris, 1963) for particle physics.

In the macroevolutionary context, the first branching model that must be mentioned is the Yule model (1924) or equivalently the Equal Rate Markov (ERM) model (Aldous, 2001a). In this model, starting from a trivial tree (a single tip: the root), a tip is chosen randomly for branching. This process is repeated until the desired number of tips attained. It is very well known that this model has a scaling $\langle d \rangle_n \sim \log n$, like the symmetric tree with a difference in coefficient. This scaling of ERM model trees is calculated at the end of this section. Also, another model with the similar scaling of ERM is added at the end. Indeed, it had been well known for long time that estimated trees' scaling is faster than $\log n$ which makes this model inappropriate in this context for the balance structure.

Other than the ERM model, there existed some earlier use of branching processes in macroevolutionary literature, like Ref (Cavalli-Sforza and Edwards, 1967). However, the approach from stochastic branching modeling of macroevolution, i.e., production of trees and studying the shape of trees, can be traced back to a group of paleontologists and population biologist at Wood Hole in the early 1970s (Mooers and Heard, 1997). Current studies (Blum and François, 2006; Brunet et al., 2008; Pinelis, 2003), including ours, of stochastic production of trees and their comparison to phylogenetic trees can be seen the modern descendants of Wood Hole school. As mentioned in other references such as (Blum and François, 2006), it is believed that the comparison of stochastic branching models to data would help to understand underlying macroevolutionary process.

To our knowledge, the only relatively successful model catching the observed balance shape is the AB model which is a subfamily of Beta-Splitting Model introduced by Davis Aldous in Ref. (Aldous, 1996). The motivation behind Aldous' 1996 study is to find a one-parameter mechanism on cladograms which can result in trees similar to the published reconstructed phylogenetic trees. The model can be described as follows: Place n species randomly on the interval $(0, 1)$. Choose a random number x from probability density function $f(x)$ and divide the interval into two parts such that i species stay in the subinterval $(0, x)$ and $n - i$ species stay in the subinterval $(x, 1)$. Having renormalized each subinterval into $(0, 1)$ and rearranged the places of species, keep going on the procedure unless a subinterval contains only one species (See Figure 1.6 for a graphical explanation). The name of "Beta" was chosen because $f(x)$ is related to a special beta function, $x^\beta(1-x)^\beta$. Aldous gives an outline of the proof of the mean depth scaling $\langle d\rangle_n \sim \log^2 n$ of the AB model (Beta Splitting model, $\beta = -1$) in Ref. (Aldous, 1996). Ref (Blum and François, 2006) gives a detailed report regarding the resemblance of TREEBASE and AB Model trees. However, a biological interpretation of the model is still missing.

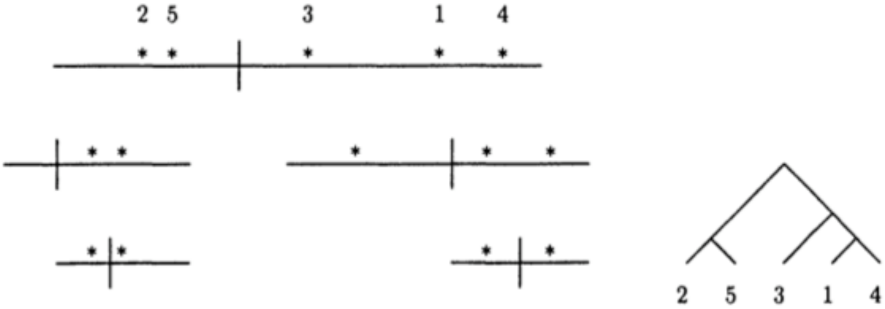

Figure 1.6: Explanation of the AB Model (Beta-Splitting Model with $\beta = -1$)(Aldous, 2001a)

Thus, we seek for stochastic branching models whose mean depth scaling is similar to the reconstructed phylogenetic trees. We should emphasize that in all our models, we build our system such that the species are basic units of dynamics. In other words, we ignore the dynamics occurring below the species level, which is the subject of microevolution. This tenet of modeling of macroevolution independent of microevolution is becoming common in the literature and according to Reznick and Ricklefs (2009) it is in parallel to Darwin's ideas on the subject.

1.6.1 Mean Depth Scaling of the ERM Model Trees

In a mean field description, the sum of depths is increased by mean depth plus 2. Let us write this,

$$n\langle d\rangle_n = (n-1)\langle d\rangle_{n-1} + \langle d\rangle_{n-1} + 2 \tag{1.7}$$

which rearranging leads us to

$$\langle d\rangle_n - \langle d\rangle_{n-1} = \frac{2}{n} \tag{1.8}$$

the L.H.S is like d/dn when $n >> 1$, so

$$\frac{d\langle d\rangle_n}{dn} = \frac{2}{n}, \tag{1.9}$$

which can be solved as

$$\langle d\rangle_n = 2\log n + const. \tag{1.10}$$

One can force the $n = 2$ and $\langle d \rangle_n = 1$ as the initial condition and get $const = 1 - 2\log 2$ and finally

$$\langle d \rangle_n = 2\log n + 1 - 2\log 2 \tag{1.11}$$

which is said to be scaling as

$$\langle d \rangle_n \sim \log n \tag{1.12}$$

Note that it is a same scaling behavior as the symmetric tree with differing in coefficient.

1.6.2 Mean Depth Scaling of Continuous time Independent Branching Model with Exponential Lifetime Distribution

Here we present other type of branching model, developed in continuous time, which could be used to model macroevolution: a continuous-time independent branching model. It turns out to give a scaling coincident with the ERM model at least when an exponential lifetime distribution, the one that will be considered here, is used.

In this model, starting at $t = 0$, a species is created and a lifetime τ is assigned to it from a probability distribution function (pdf) $f(\tau)$. This is the function such that the probability of having a lifetime between τ and $\tau + d\tau$ is $f(\tau)d\tau$. After τ time passes a binary branching occurs from this species and the same procedure is repeated for each newly created species. This process will proceed until the desired number of existing species n or total time t is reached in the growing model.

Let us introduce $F(\tau)$, cumulative pdf of $f(\tau)$, i.e. $\int_0^\tau f(\tau')d\tau'$. It refers to the probability to assign a life time smaller than τ. In our model, both f and F are 0 for $\tau = 0$; $Q(\tau)$ is the probability to assign a life time bigger than τ. It is simply $1 - F(\tau)$ or $\int_\tau^\infty f(\tau')d\tau'$; $r(\tau)$ is another way of describing the probability that a branching occurs in the interval τ and $\tau + d\tau$ can be expressed by a rate $r(\tau)d\tau$ where not branching becomes $1 - r(\tau)d\tau$. We can relate the $r(\tau)$ to $f(\tau)$ as follows: Consider $n \to \infty$, $\Delta\tau \to 0$ and $N\Delta\tau = \tau$, then we can immediately write

$$Q(\tau) = \lim_{\Delta\tau \to 0} \prod_{i=0}^{n} (1 - r(\tau_i)\Delta\tau). \tag{1.13}$$

Remembering $e^{-x} = 1 - x + ...$ for small x's we can write that

$$Q(\tau) = e^{-\int_0^\tau r(\tau_i')d\tau'}. \tag{1.14}$$

Taking logarithm of both sides and differentiating with respect to τ gives

$$\frac{d}{dt} \log Q(\tau) = -r(\tau), \tag{1.15}$$

and equivalently we write $r(\tau)$ in terms of $Q(\tau)$ as

$$r(\tau) = -\frac{\dot{Q}(\tau)}{Q(\tau)}. \tag{1.16}$$

Now, it is easy to find $f(\tau)$ in terms of $r(\tau)$ by remembering that there is no event occurred until that moment AND exactly one event occurred at that moment:

$$f(\tau) = Q(\tau)r(\tau) \tag{1.17}$$

$$f(\tau) = r(\tau)e^{-\int_0^\tau r(\tau_i')d\tau'} \tag{1.18}$$

Using above expression (1.18) it easy to see that an equal (constant, say $r(t) = \lambda$) rate would yield an **exponential** decreasing life time distribution, i.e. $f(\tau) = \lambda e^{-\lambda\tau}$.

For a given $f(\tau)$, what is the scaling of the mean depth with number of species? In order to seek an analytical answer to this question at first we will do followings. Firstly, we will derive time dependence of the number of existing species $n(t)$ and the sum of depths $D(t)$. Then, assuming $\langle d\rangle = \langle D/n\rangle = \langle D\rangle/\langle n\rangle$, we will obtain the mean depth scaling with respect to n.

For $n(t)$, let us consider a root species with a lifetime τ and its branching into 2 subtrees. Since the model does not have any memory and no interaction between any nodes we are able to write the following convolution:

$$N(t) = Q(t) + 2\int_0^t N(t-\tau)f(\tau)d\tau, \tag{1.19}$$

where N is the total number of nodes in tree (remember again $N = 2n-1$). Writing in terms of $n(t)$,

$$2n(t) - 1 = Q(t) + 4\int_0^t n(t-\tau)f(\tau)d\tau - 2F(t). \tag{1.20}$$

For $D(t)$ we consider two statistically the identical and uncorrelated trees and unite them by creation of a root species with τ lifetime similar to the above case. Then one can write

$$D(t) = 2\int_0^t [D(t-\tau) + n(t-\tau)]f(\tau)d\tau \tag{1.21}$$

One strategy to solve the above convolution expressions is to transform them into Laplace space. Therefore, Eq.(1.20) becomes

$$2\tilde{n}(s) - 1/s = \tilde{Q}(s) + 4\tilde{n}(s)\tilde{f}(s) - 2\tilde{F}(s) \tag{1.22}$$

Remembering the relations between Q and F with f, we arrive at equation

$$\tilde{n}(s) = \frac{1}{s}\frac{2-3\tilde{f}(s)}{2-4\tilde{f}(s)} \tag{1.23}$$

Similarly dealing with Eq.(1.21), we write the following equation,

$$\tilde{D}(s) = \frac{2\tilde{n}(s)\tilde{f}(s)}{1-2\tilde{f}(s)}, \tag{1.24}$$

and substituting Eq.(1.23) we arrive at equation,

$$\tilde{D}(s) = \frac{1}{s}\frac{\tilde{f}(s)(2-3\tilde{f}(s))}{(1-2\tilde{f}(s))^2}. \tag{1.25}$$

Therefore, for given $f(\tau)$ (and so $\tilde{f}(s)$) we might try to seek for inverse Laplace transfroms of $\tilde{n}(s)$ in Eq.(1.23) and $\tilde{D}(s)$ in Eq.(1.25) and obtain mean depth as $\langle d\rangle(t) = D(t)/n(t)$.

Suppose we have $f(\tau) = \lambda e^{-\lambda\tau}$ and so $\tilde{f}(s) = \lambda/(s+\lambda)$, then according to Eq.(1.23),

$$\tilde{n}(s) = \frac{1}{s}\frac{s-5\lambda/2}{s-3\lambda} \tag{1.26}$$

that can be written as

$$\tilde{n}(s) = \frac{1}{s} + \frac{1}{s}\left(\frac{\lambda/2}{s-3\lambda}\right) \tag{1.27}$$

the inverse Laplace transform of the above equation can be found by recalling the theorem $L^{-1}(\tilde{f}(s)/s) = \int_0^t f(u)du$ (Spigel, 1970). Therefore,

$$n(t) = 5/6 + \lambda/6e^{3\lambda t}. \tag{1.28}$$

Similar treatment for solving the Eq.(1.25)

$$\tilde{D}(s) = \frac{1}{s}2\lambda\Big\{\frac{1}{s-3\lambda} + \frac{1/2\lambda}{(s-3\lambda)^2}\Big\}, \tag{1.29}$$

$$D(t) = \int_0^t 2\lambda\Bigg\{e^{3\lambda t'} + \lambda/2e^{3\lambda t'}t'\Bigg\}dt', \tag{1.30}$$

which result in

$$D(t) = 2\lambda\Bigg\{\frac{e^{3\lambda t}-1}{3\lambda} + \lambda/2\Big\{\frac{te^{3\lambda t}}{3\lambda} + \frac{1-e^{3\lambda t}}{(3\lambda)^2}\Big\}\Bigg\}, \tag{1.31}$$

and putting in a clearer way:

$$D(t) = 3\lambda t e^{3\lambda t} + 5e^{3\lambda t} - 5, \tag{1.32}$$

dividing $D(t)$ with $n(t)$ we have expected mean depth $d(t)$:

$$\langle d\rangle(t) = 3t + 5/\lambda + ... \tag{1.33}$$

For big λt values we have $t \sim \log(n)$, from Eq.(1.28). Therefore, if we put this into the last expression we have, we get logarithmic scaling behavior for the mean depth, i.e.,

$$\langle d\rangle_n \sim \log(n). \tag{1.34}$$

2

Summary of the Original Research

Our general **motivation** in this thesis research is the elucidation of the underlying laws governing macroevolution. Particularly, we are interested in understanding how speciation occurs and what diversification patterns result from it. In other words, we are questioning how the structure of branching events in the hypothetical tree of life is. At present, the phylogenetic trees, objects of evolutionary relationships of extant species, serve as estimated diversification patterns of macroevolution which we intend to appreciate.

One approach to understanding the particular shape of these speciation branchings is via comparing them with the ones produced by well-known (proposed) branching mechanisms. Whenever a satisfying similarity at the structure is observed, the rules producing the tree might be interpreted biologically. This can lead to theories of macroevolution, such as a neutral theory of macroevolution which is sought after the success of the neutral theory of molecular evolution. To our knowledge, there does not exist yet any particular branching model which both captures the structure of the phylogenetic trees and is interpreted as meaningful in biological sense.

Therefore, as the main **aim** of our research, we seek biologically inspired models of growing trees that exhibit similar topology with estimated trees found in databases.

Here, we present Chapters 3 and 4, which are the contents of two research articles, and which include three simple branching models, named as *Ford's Alpha*, *Activity* and *Age* models. We use the balance notion, i.e., left-right symmetry, for topological characterization of the trees. We quantify it by using the *mean depth* defined as

$$\langle d \rangle_n = (\sum_{i=1}^{n} d_i)/n,$$

where d_i is the number of ancestors of tip i and n is the total number of tips in tree. We are particularly interested in the behavior of the mean depth scaling in order to compare the ensembles of trees from models and/or databases. After the mean depth scaling analysis of the TREEBASE (species phylogeny) which takes place in both of the articles and of the PANDIT (protein phylogeny) which takes place in the second article, we conclude the following reference scaling information,

$$\langle d\rangle_n \sim \begin{cases} n & \text{for the most imbalanced tree} \\ \log n & \text{for the most balanced, ERM model trees} \\ (\log n)^q & \text{for TREEBASE, PANDIT trees} \end{cases}$$

where q is clearly greater than 1, with the best fit to $q \sim 2$ in good agreement with some other reports claiming $q = 2$.

The first model presented in Chapter 3 is the **Ford's Alpha Model** which was proposed by D. J. Ford. It is a one-parameter (α) family tree growing model where $0 < \alpha < 1$. At each branching event, which does not necessarily happens at the tips in this model, an edge is chosen for adding a new edge and a new tip, where beforehand a branching probability proportional to $1 - \alpha$ and α is assigned to each leaf and internal edge, respectively. The mean depth scaling property was studied and expressed analytically by Ford as

$$\langle d\rangle_n \sim \begin{cases} n^\alpha & 0 < \alpha \leq 1 \\ \log n & \alpha = 0 \end{cases}$$

Our numerical studies have verified these results if n is large enough. Although this model gives a simple mechanism for scaling with one parameter, its interpretation in evolutionary sense is hard to justify unless one comments it in the context of the error arising of phylogenetic methods.

The second model presented in Chapter 3 is the **Activity Model**. It is a one-parameter (p) family tree growing model proposed by us where $0 < p < 1$. Starting from a trivial tree, at each step, one active tip is chosen to branch into two new tips. Each new tip is assigned to be active or inactive with a probability proportional to p and $1 - p$, respectively. We show both analytically and numerically mean depth scaling as

$$\langle d\rangle_n \sim \begin{cases} n^{1/2} & p = 1/2 \\ \log n & \text{otherwise.} \end{cases}$$

The activity model again does not have very clear biological meaning. However, it has the mechanism of the birth-death critical branching within a framework of transitions between node internal states which can help identify the proper biological meaning with an additional study.

The last model we hold here is presented in Chapter 4 and named as **Age Model**. It is one-parameter (α) family tree growing model proposed by us where $-\infty < \alpha < \infty$. It is a discrete time incremented model where discrete step Δt is considered to be 1 and $1/n$ as two different versions. At a given time, each tip i possesses an age $\tau_i(t)$ which is the time passed from the birth of the tip, t_i, to present time t, i.e. $\tau_i = t - t_i$. Two new tips, are added to the tip i after each time increment Δt with a probability proportional to $\tau_i^{-\alpha}$.

Numerically, we show that for $\alpha = 1$, both cases show a mean depth scaling $\langle d \rangle_n \sim (\log n)^2$ which is claimed that database trees display, too. Also, we observed that the mean depth scaling goes rapidly to that of the comb tree ($\sim n$) and of the symmetric tree ($\sim \log n$) for $\alpha > 1$ and $\alpha < 1$, respectively. Analytically investigating the $\Delta t = 1$ case, we conclude the following mean depth scaling behavior:

$$\langle d \rangle_n \sim \begin{cases} \log n & \text{if } \alpha < 1 \\ (\log n)^2 & \text{if } \alpha = 1 \\ n^{\alpha - 1} & \text{if } 1 < \alpha < 2 \\ \int dn / \log n & \text{if } \alpha = 2 \\ n & \text{if } \alpha > 2 \end{cases}$$

Thus, at a critical parameter, $\alpha = 1$, the Age model produces similar trees to the reconstructed evolutionary trees. It matches the well known observation that the species which are coming from evolutionary lines of many diversification, did (will) tend to diversify more than those do not. If an age property for species is found, then it might be used as a null model (theory) for macroevolution. Our research on this model and its interpretation are still continuing.

Part II

Original Research

3

Simple models for scaling in phylogenetic trees

Abstract[1]: Many processes and models –in biological, physical, social, and other contexts– produce trees whose depth scales logarithmically with the number of leaves. Phylogenetic trees, describing the evolutionary relationships between biological species, are examples of trees for which such scaling is not observed. With this motivation, we analyze numerically two branching models leading to non-logarithmic scaling of the depth with the number of leaves. For Ford's alpha model, although a power-law scaling of the depth with tree size was established analytically, our numerical results illustrate that the asymptotic regime is approached only at very large tree sizes. We introduce here a new model, the *activity* model, showing analytically and numerically that it also displays a power-law scaling of the depth with tree size at a critical parameter value.

3.1 Phylogenetic branching and models

Although most modern studies on complex networks (Albert and Barabási, 2002; Boccaletti et al., 2006) consider situations in which nodes are connected by multiple paths, the case of *trees*, i.e. graphs without closed cycles, is relevant to describe many natural and artificial systems. Branching in real trees (Stevens, 1974), in blood vessels (West et al., 1997), in river networks (Rodriguez-Iturbe and Rinaldo, 1997) or in computer file systems (Klemm et al., 2005, 2006) produce complex tree patterns worth to be described and understood. Trees are the outcome of classifications algorithms (Jain and Dubes, 1988) and of branching processes (Harris, 1963) and they also arise when computing community structure (Guimerà et al., 2003) or as a backbone (for example a minimum spanning tree) for more connected networks (Garlaschelli et al., 2003; Hernández-García et al., 2007; Rozenfeld et al., 2008).

[1] Emilio Hernández-García, **Murat Tuğrul**, E. Alejandro Herrada, Víctor M. Eguíluz and Konstantin Klemm, (2009), accepted to International Journal of Bifurcation and Chaos.

Evolutionary processes leading to speciation are also summarized in phylogenetic trees (Cracraft and Donoghue, 2004). In these trees the leaves represent living species and each internal node represents a branching event in which an ancestral species diversified into daughter species. Every internal node is thus the root of its associated subtree which consists of all its descendant nodes. Phylogenetic tree topology encodes information on evolutionary mechanisms which is beginning to be scrutinized (Blum and François, 2006; Burlando, 1990, 1993; Ford, 2006; Hernández-García et al., 2007; Herrada et al., 2008).

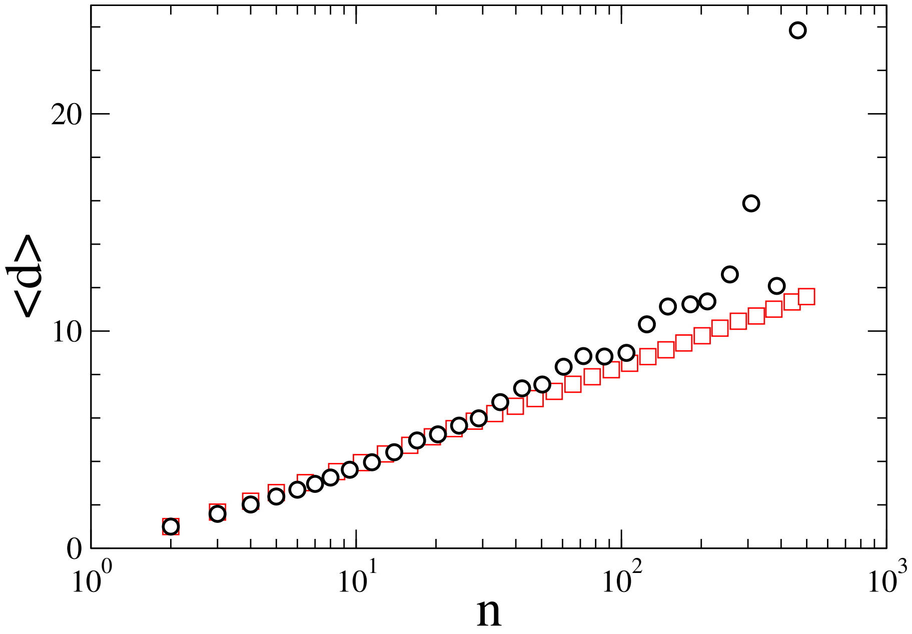

Figure 3.1. Mean depth $\langle d \rangle$ of trees in TreeBASE (circles) as a function of number of leaves n. Squares are obtained from computer simulations of the ERM model, behaving as Eq. (3.1) for large n. At large sizes, the depth in the real phylogenetic trees scales with the number of leaves faster than the ERM behavior. For both real phylogenies and model, depth values for each tree size are obtained by logarithmic binning of the depth of all trees and subtrees with that size.

The earliest mathematical model of evolutionary branching was proposed by Yule (1925). Apart from the distinction he introduced between genera and species diversification, the model is equivalent to the Equal Rates Markov (ERM) model (Cavalli-Sforza and Edwards, 1967; Harding, 1971): starting from a single ancestral species, one among the tree leaves existing at the present time is chosen at random, bifurcating into two new leaves. Then this operation is repeated for a number of time steps or, equivalently, until the tree reaches a desired size. The topological characteristics of the constructed trees are surprisingly robust, being shared by apparently different models such as the coalescent and others (Aldous, 2001b). Essentially what is needed is that different branches at a given time branch independently and with the same probabilities. When extinction is taken into account, the same topology is recovered when considering only the lineages surviving at the final time. One of the

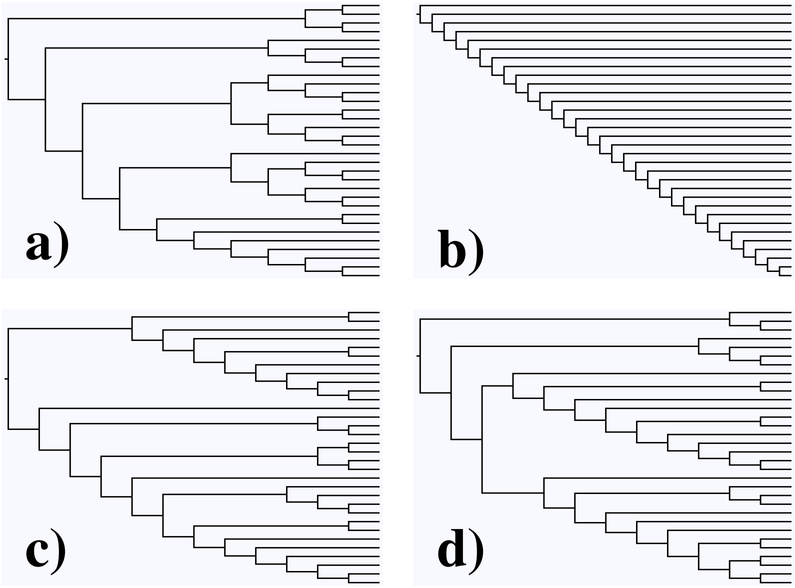

Figure 3.2. Examples of trees with 32 leaves, generated from several models. a) Tree generated with the ERM model, which is equivalent to the alpha model with $\alpha = 0$. b) The completely unbalanced tree, which is equivalent to the alpha model with $\alpha = 1$. c) A tree generated with the alpha model for $\alpha = 0.5$. d) A tree generated with the activity model for $p = 0.5$. The trees in c) and d) display an imbalance intermediate between a) and b).

characteristics of this type of branching is a distribution of subtree sizes A scaling at large sizes as A^{-2}, an outcome robustly observed in many natural and artificial systems and in classification schemes, including taxonomies (Burlando, 1990; Caldarelli et al., 2004; Capocci et al., 2008). Another important characteristic is that the mean depth of the tree $\langle d \rangle$ (i.e. the average distance, measured in number of links, from the leaves to the root) scales logarithmically with the number of leaves n:

$$\langle d \rangle \sim \log n \ . \tag{3.1}$$

It is worth noting that these results apply not only to many random branching models, but also to the simple deterministic Cayley tree, in which all internal nodes at a given level split in a fixed number of daughter nodes.

In view of this generality it was surprising to find that the topology of observed phylogenies does not agree with any of these predictions (Herrada et al., 2008). In fact, it was known since some time ago that real phylogenies are substantially more *unbalanced* than predicted by the ERM and similar models (Aldous, 2001b; Blum and François, 2006). This means that some lineages diversify much more than others, in a way that is statistically incompatible with the ERM predictions. Figure 3.1 compares data (Herrada et al., 2008) compiled from TreeBASE, a public repository containing several thousands of empirical phylogenetic trees corresponding to virtually all kinds of organisms in Earth, with the predictions of the ERM model. For the phylogenetic trees at large sizes the mean depth scales with the number of leaves faster than the ERM behavior in Eq.(3.1).

The breakdown of the ERM behavior indicates that evolutionary branching should present correlations either in time or between the different branches. Mechanisms producing trees with non-ERM scaling for the depth have been identified, as for example the situation of critical branching (De Los Rios, 2001; Harris, 1963) or optimization of transport processes (Banavar et al., 1999). In the phylogenetic context models of this type have been proposed (Aldous, 2001b; Blum and François, 2006; Ford, 2006; Pinelis, 2003), although most of them lack a clear interpretation in biological terms.

In the following we present results for two branching models showing asymptotically non-ERM, i.e. non-logarithmic, scaling for the depth. Their study is motivated, on the one hand, by the empirical results above from real phylogenetic trees. On the other, they pertain to the small set of available models with non-ERM scaling which are defined *dynamically* (i.e. by a set of rules that are applied to the present state of a growing tree to find the state at the next time step) rather than being characterized globally by statistical or optimization prescriptions. The first model we present, Ford's *alpha* model, is a simple example for which the non-trivial asymptotic scaling (of the power law type) has been analytically identified. We analyze it numerically to confirm this prediction and to display the behavior at finite sizes. We introduce later a new model, the *activity* model, which also leads to non-logarithmic depth scaling at a critical parameter value.

3.2 Ford's alpha model

Ford (2006) introduced a model for recursive tree formation: At a given step in the process the tree is a set of leaves connected by terminal links to internal nodes, which are themselves connected by internal edges until reaching the root (the root itself is considered to have a single edge, which we count as internal, joining to the first bifurcating internal node; with this convention a tree of n leaves has $n-1$ internal edges). Then, a probability of branching proportional to $1-\alpha$ is assigned to each leaf, and proportional to α to each internal edge. By normalization these probabilities are, respectively, $(1-\alpha)/(n-\alpha)$, and $\alpha/(n-\alpha)$. When a leaf is selected for branching, it gives birth to a couple of new ones, as in the ERM model. But when choosing an internal edge, a new leaf branches from it by the insertion in the edge of a new internal node. For $\alpha = 0$ we have the standard ERM model. For $\alpha = 1$ the completely unbalanced comb tree, in which all leaves branch successively from a main branch, is generated. Intermediate topologies are obtained for $\alpha \in (0, 1)$. Figure 3.2 shows examples of trees generated for different values of α.

By considering the effect of the addition of new leaves on the distances between root and other nodes, Ford (2006) derived exact recurrence relationships which, when

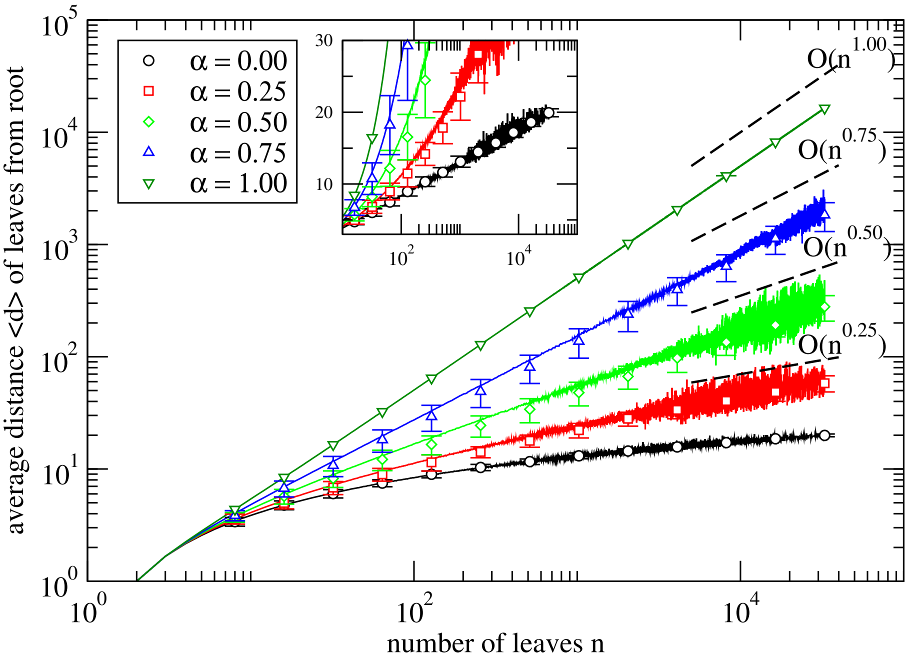

Figure 3.3. Depth statistics vs tree size for the alpha model. Symbols indicate the mean depth of leaves from root, averaged over the 100 trees generated for each size (2^k, $k = 3, 4, ..., 15$), and the error bars are the corresponding standard deviations. The points in the rugged lines come from each subtree of all trees generated. The dashed segments indicate the analytic predictions (Ford, 2006) for the scaling at large n. The inset highlights the logarithmic scaling of the $\alpha = 0$ case.

written in terms of the average depth, lead to:

$$\langle d \rangle_{n+1} = \frac{n}{n-\alpha} \langle d \rangle_n + \frac{2n(1-2\alpha)}{(n+1)(n-\alpha)} \ . \tag{3.2}$$

$\langle d \rangle_n$ is the mean depth of the leaves of a tree with n leaves. By assuming a behavior $\langle d \rangle_n \sim n^\nu$ at large n, and expanding Eq. (3.2) in powers of $1/n$, we get $\nu = \alpha$, so that

$$\langle d \rangle_n \sim n^\alpha \text{ ,if } \ 0 < \alpha \leq 1 \ . \tag{3.3}$$

If $\alpha = 0$ the standard ERM behavior, Eq. (3.1), is recovered.

Figure 3.3 shows numerical results for the depth of trees generated with this model. Note that the predicted asymptotic behavior is attained but only at very large tree sizes, in general sizes much larger than the tree sizes of the examples shown in Fig. 3.2 and of the available empirical phylogenies. As analytically demonstrated (Ford, 2006) depth statistics of subtrees of given size extracted from a large tree behave as data from trees of that size directly generated by the alpha model algorithm.

While the Ford model gives a simple mechanism for scaling in trees with a tunable exponent, the dynamical rule of posterior insertion of inner nodes is hard to justify in the context of evolution (although one can think on the modelling of errors arising in phylogenetic reconstruction methods when incorrectly assigning a splitting to a non-existing ancestral species). This motivates the introduction of a new model described in the next section.

3.3 Activity model

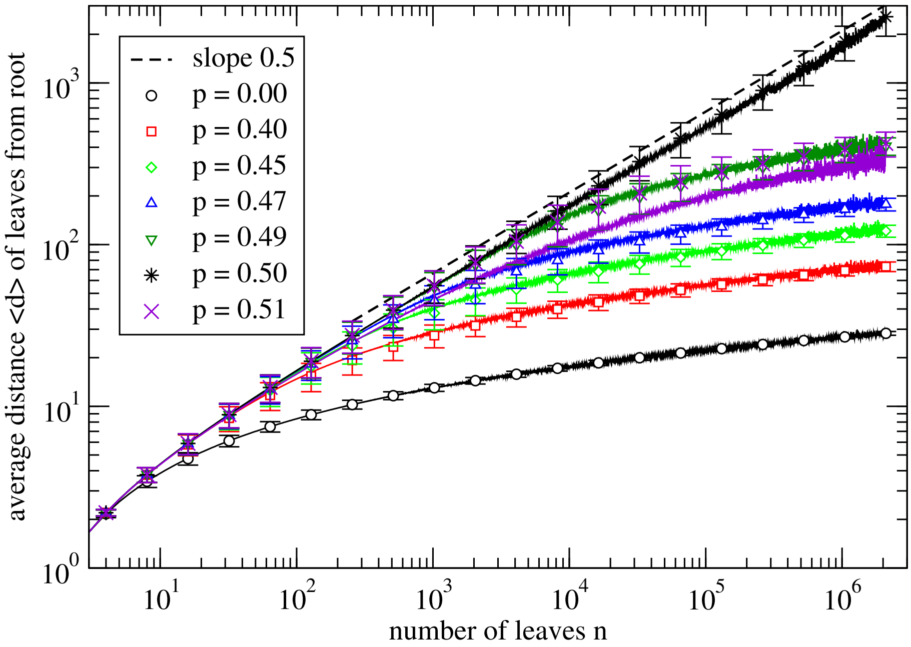

Figure 3.4. Average depth versus size for the activity model for various values of the activation probability p. Data points displayed by symbols give the average distance of leaves with respect to the root. Error bars give the standard deviation taken over different realizations (1000 trees per data point). Data in the rugged curves are for all subtrees of trees with size $2^{21} = 2097152$. The dashed line represents a power law scaling with exponent 1/2, corresponding to the scaling of the $p = 0.5$ curve, as discussed in the text.

In this section we show that tree shapes distinct from the ERM model may also result from a memory in terms of internal states of the nodes. The *activity* model proposed here is conceptually similar to the class of models suggested by (Pinelis, 2003). However, the present model distinguishes only between active and inactive nodes and has a single parameter controlling the spread of activity.

Starting from a single node (the root), a binary tree is generated as follows. At each step, a leaf i of the tree is chosen and branched into two new leaves. Each of the two new leaves, independently of the other, is set active with probability p or inactive with probability $1-p$. The branching leaf i is chosen at random from the set of active leaves if this set is non-empty. Otherwise, i is chosen at random from the set of all leaves. Figure 3.4 shows that for $p = 1/2$ the model generates trees with mean depth growing as the square root of tree size (note the log-log scale). Figure 3.2 displays a small-size example of such trees. For values of p below or above 1/2, $\langle d \rangle$ seems to increase logarithmically with n.

Here we give a simplified argument to understand the observed exponent 1/2 of the distance scaling with system size in the case $p = 1/2$. At the time the growing tree has n leaves in total, let $D_a(n)$ be the expected sum of distances of active leaves from the root, and $D_b(n)$ the analogous quantity for the inactive leaves. When a randomly chosen active leaf –at distance d_a from root– branches, the expected increase of $D_a(n)$ is

$$\begin{aligned}\Delta D_a(n) &\equiv D_a(n+1) - D_a(n) = \\ p^2(d_a+2) &+ 2p(1-p)\cdot 1 + (1-p)^2(-d_a) \\ &= (2p-1)d_a + 2p \ . \end{aligned} \tag{3.4}$$

Here the three terms of the second line are for the activation of two, one and zero of the new leaves, respectively. This expression is appropriate as far as the number of active nodes is not zero. Simultaneously, the expected change in $D_b(n)$ during the same event is

$$\begin{aligned}\Delta D_b(n) &= \\ p^2 \cdot 0 &+ 2p(1-p)(d_a+1) + (1-p)^2 2(d_a+1) \\ &= 2(1-p)(d_a+1) \ . \end{aligned} \tag{3.5}$$

We now average $\Delta D_a(n)$ over the different choices of the particular active leave that has been branched. This amounts to replacing d_a in the above formulae by $\langle d_a \rangle_n$, the average depth of the *active* leaves in a tree of n leaves. Writing $D_i(n+1) = D_i(n) + \Delta D_i(n)$, for $i = a, b$, one would get a closed system for the quantities $D_i(n)$ provided $\langle d_a \rangle_n$ is expressed in terms of them. This can be done by writing $\langle d_a \rangle_n = D_a(n)/a(n)$, where $a(n)$ is the expected number of active leaves in a tree of n leaves. This expected value is used here as an approximation to the actual number of active leaves.

The recurrence equations for $D_i(n)$ are specially simple in the most interesting case $p = 1/2$, since the dependence in $\langle d_a \rangle_n$ disappears from one of the equations:

$$D_a(n+1) = D_a(n) + 1 \tag{3.6}$$

$$D_b(n+1) = D_b(n) + \langle d_a \rangle_n + 1 \ . \tag{3.7}$$

The solution (with initial condition $D_a(1) = 0$) of Eq. (3.6) is simply:

$$D_a(n) = n - 1 \ . \tag{3.8}$$

Since the probabilities of an increment or decrement (by one unit) of the number of active leaves are the same and time-independent for $p = 1/2$, the number of active nodes performs a symmetric random walk with a reflecting boundary at 0 (this last condition arises from the prescription of setting active one node when the number of active nodes has reached zero in the previous step). For such random walk the expected value of active leaves $a(n)$ increases as the square root of the number of steps. Since a new leaf is added at each time step, this leads to:

$$a(n) \sim n^{1/2} \ . \tag{3.9}$$

Combining (3.8) and (3.9) we obtain the average distance of active nodes from root at large tree sizes:

$$\langle d_a \rangle_n \approx \frac{D_a(n)}{a(n)} \sim n^{1/2} \ . \tag{3.10}$$

Now we can plug this result into Eq. (3.7), which can be solved recursively:

$$D_b(n) = D_b(1) + \sum_{t=1}^{n-1} (\langle d_a \rangle_t + 1) \sim \sum_{t=1}^{n-1} t^{1/2} \sim n^{3/2} \ . \tag{3.11}$$

The totally averaged depth $\langle d \rangle_n$, which counts both the active and the inactive leaves, is

$$\langle d \rangle_n = \frac{D_a(n) + D_b(n)}{n} \sim \frac{n^{1/2} + n^{3/2}}{n} \sim n^{1/2} \ , \tag{3.12}$$

which explains the asymptotic behavior observed in Fig. 3.4 for $p = 1/2$.

We note that the growth dynamics presented here may be mapped to a branching process (Harris, 1963), with the difference that here the death (inactivation) of a node does not lead to its removal from the tree. The special case $p = 1/2$ corresponds to a critical branching process.

3.4 Discussion

We have presented and studied two simple models which lead to non-logarithmic scaling of the tree depth. In contrast with many of the available models having this behavior (Aldous, 2001b; Banavar et al., 1999; Blum and François, 2006; Ford, 2006) they are formulated as *dynamical* models involving *growing trees*, so that rules are given to obtain the tree at the next time step from the present state. Their study has been motivated by data from phylogenetic branching, and they are interesting additions to our present understanding of complex networks and trees.

A recent analysis of several evolutionary models including species competition (Stich and Manrubia, 2008) indicates that in these models correlations are finally destroyed by mutation processes and persist only for a finite correlation time. Thus sufficiently large trees would have a scaling behavior closer to the asymptotic ERM predictions. Since the largest phylogenies in databases such as TreeBASE have only some hundreds of leaves, it is possible that the observed imbalance and depth scaling is a finite-size regime. Nevertheless models going beyond the ERM scaling are needed at least to explain this finite-size regime, and also to elucidate the true asymptotic scaling behavior. Here, we have also observed large finite-size transients in the alpha model of Sect. 2.

The different types of scaling of depth with size can be interpreted as indicating different values of the (fractal) dimensionality of the trees. This is so because $\langle d \rangle$ is a measure of the *diameter of the tree*, and because for a binary tree the total number of nodes is simply twice the number of leaves. Since the simplest definition of dimension D of a network (Eguíluz et al., 2003) is given by the growth of the number of nodes as the diameter increases, $n \sim \langle d \rangle^D$, power law scaling of the type $\langle d \rangle \sim n^\nu$ indicates that the tree can be thought as having a dimension $D = 1/\nu$. The logarithmic scaling in the ERM model is an example of the *small-world* behavior common to many network structures (Albert and Barabási, 2002), which is equivalent to having an effective infinite dimensionality, whereas the power law scaling reveals a finite dimension for the tree, which implies a more constrained mode of branching. The alpha model produces trees with tunable dimension from 1 to ∞, and the critical activity model gives two-dimensional trees.

The final aim of the modelling of phylogenetic trees is to provide biological mechanisms explaining the branching topology of the Tree of Life. In this direction, the branching of internal edges in the Ford model has no obvious biological interpretation. The activity model puts the mechanisms of birth-death critical branching (Harris, 1963) within a framework of transitions between node internal states similar in spirit to the approach of (Pinelis, 2003). The need to tune a parameter to attain the non-ERM critical behavior is however a limitation for its applicability. Much additional work is needed to identify the proper biological mechanisms behind evolutionary branching and adequate modelling of them.

Acknowledgments

We acknowledge financial support from the European Commission through the NEST-Complexity project EDEN (043251) and from MICINN (Spain) and FEDER through project FISICOS (FIS2007-60327).

4

Might an Age Dependent Branching Model Help Us to Understand Macroevolution?

Abstract[1]**:** Understanding the rules of macroevolution remains a fundamental topic of biological research. Phylogenetic trees serve as an estimated macroevolutionary relationship structure, usually obtained directly from molecular data. In this work, we first analyse TREEBASE and PANDIT databases for tree imbalance and find scaling behavior with tip (species) number n, which has not been explained yet by branching models with biological interpretation. We then propose a one-parameter family of branching models in which the branching probability depends on the age of the species. We investigate it analytically and computationally, exploring the transitions among different behaviors. We identify a member of the family producing imbalance scaling as $\log^2 n$, in good agreement with the behavior found in the databases.

4.1 Introduction

The evolution of life has been one of the most intriguing fields in science at least since the publication of Darwin's *On the Origin of Species* (Darwin, 1859). Traditionally, the research field of biological evolution has made a sharp distinction between microevolution and macroevolution (Erwin, 2000). Microevolution deals with changes occurring within a species or small group of organisms in comparably short time. Availability of vast empirical datasets has allowed extensive and systematic research of microevolutionary processes. Macroevolution is the long-term dynamics of the set of species. Species cease to exist (extinction) when the reproduction is inhibited by environmental pressure. New species come into existence by diversification

[1] The content of this chapter are part of **Murat Tuğrul**, Stephanie Keller-Schmidt, Víctor M. Eguíluz, Emilio Hernández-García and Konstantin Klemm, *work in progress*

(speciation) of existing ones.

Following the concept of vertical evolution, the speciation throughout the history can be presented as a tree where tips are extant species and internal nodes are speciation events. The real tree representing the history of life is not observable itself, except for some estimation from fossil data which is open to scientific controversies. However, phylogenetic methods reconstruct trees of evolutionary relationship between a number of present-day species, proteins, etc. by using molecular data from them. The resulting trees are generally binary trees and named *phylogenetic trees*. Scrutinizing these structures is useful to elucidate properties of macroevolution (Barraclough and Nee, 2001; Reznick and Ricklefs, 2009).

From the mathematical point of view, a tree is a graph (or network) without cycles. One of the notions characterizing the topology of a tree is its balance, meaning its degree of left-right symmetry.

There are many databases of such reconstructed (estimated) evolutionary trees available today and the balance of the trees have been analysed in the literature (Blum and François, 2006; Herrada et al., 2008). To our knowledge, the balance characteristics of phylogenetic trees have not been explained yet by a biologically inferred model and thus remains as a gap in evolution research.

It is important to note that there are discussions on the possible non-biological effects contained in reconstruction results (i.e. artifacts from estimation techniques, incompleteness of the trees, decisions of phylogeneticists, etc.; see references (Mooers and Heard, 1997) or (Barraclough and Nee, 2001) for a detailed discussion). However, we leave these discussions out of this paper and consider the observed balance scaling as a property of the diversification structure throughout the evolutionary history on earth.

Therefore, we seek biologically inspired models of growing trees that exhibit a similar balance scaling for both finite and asymptotic sizes (assuming that this scaling is universal, i.e. occurring at all tree sizes, even very large ones). In particular, we propose an age dependent branching model (hereafter called Age Model). We investigate the model by both numerical and analytical means and show that it reproduces the balance scaling in observed phylogenetic trees for a particular parameter value. In the concluding section we discuss age dependent branching as a novel interpretation of the shapes of phylogenetic trees and as potential principle of macroevolution.

4.2 Characterizing Tree Balance

Several indices for balance measurement have been proposed, used and compared in the literature (see References (Agapow and Purvis, 2002; Matsen, 2006; Mooers and Heard, 1997) for detailed discussion). After the reported comparison of the indices by (Agapow and Purvis, 2002; Matsen, 2006), considering the possibility of use in

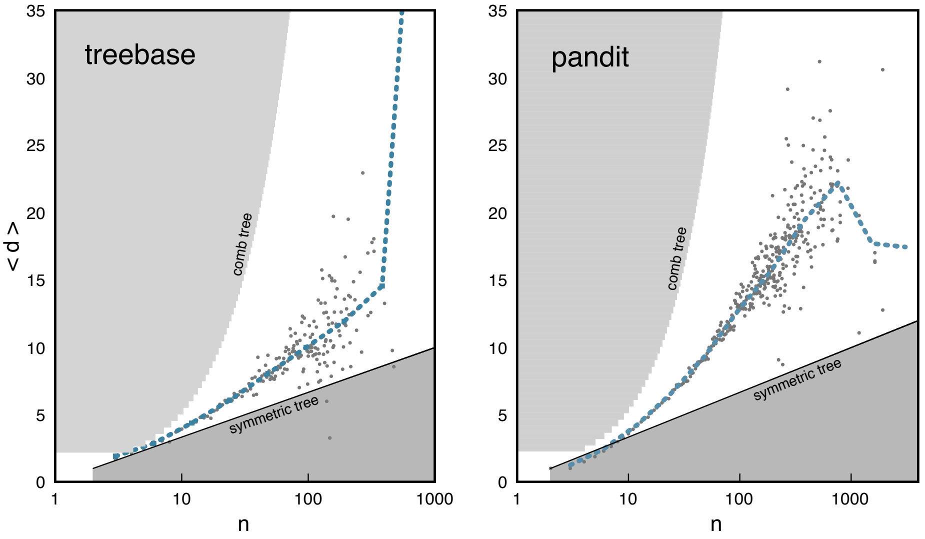

Figure 4.1. The mean depth vs. size of phylogenetic trees contained in databases of species (TREEBASE) and proteins (PANDIT) (black dots) displayed. The results are also binned in powers of 2 and averaged as shown as a blue dashed line. In this scale, the behavior $\langle d \rangle \sim \log n$ would be a straight line. Note that there are points under the symmetric tree line, which corresponds to the minimum possible depth for binary trees. This is so because the displayed symmetric tree line is for binary trees, and some polytomies occur in the data sets (more apparent in PANDIT).

polytomies, and most importantly, for the sake of a clear biological meaning and an easy-to-derive analytical tool for models, we analyse here the *mean depth* of the tree, or $\langle d \rangle$ (Sackin, 1972) being *the average number of ancestor nodes from tips to the root*, i.e.

$$\langle d \rangle = (\sum_{i=1}^{n} d_i)/n \tag{4.1}$$

where d_i is the depth, i.e., the number of ancestors, of tip i, and n is the total number of tips in tree.

In the present work we reexamined the phylogenetic databases TREEBASE (containing species phylogenies) and PANDIT (protein phylogenies) by using $\langle d \rangle$ as seen in Figure 4.1. The mean depth scaling is clearly not logarithmic, which is the outcome of the Equal Rate Model (ERM) or from a completely symmetric tree. It is neither linear as occurring for perfectly imbalanced trees. It displays instead a behavior of the type $\langle d \rangle \sim (\log n)^q$ with $q \sim 2$.

4.3 The Age Model

We introduce here a model to construct a binary tree starting from a single node, the root. Time is considered to be incremented at discrete steps of duration Δt. At a given time, each tip or leaf i is assigned an age $\tau_i(t)$ which is the time passed from the birth of the tip, t_i, to present time t, i.e. $\tau_i = t - t_i$. Two new tips, representing a pair of new species arising in a speciation event from an existing leave, i, are added to the tree after each time increment Δt. The particular tip i suffering this speciation event is chosen with probability proportional to the inverse of its age raised to a power, α which is a parameter of the model:

$$p_i(t) = \frac{c(t)}{(\tau_i)^\alpha}, \ \alpha \in \ R. \tag{4.2}$$

$c(t)$ is the normalization constant:

$$c(t)^{-1} \equiv \sum_{i=1}^{n} (\tau_i)^{-\alpha}, \tag{4.3}$$

which depends on time and, in principle, also on the particular stochastic set of ages $\tau_1, ..., \tau_n$ assigned to the tips of the tree, of size n, grown up to time t. Since we allow α to be positive and negative, the probability of choosing a tip may be both a decreasing or an increasing function of its age, but it will be shown that the empirical results are recovered for a particular positive value of α, namely $\alpha = 1$, which means a branching probability decreasing with age. Note that if $\alpha \to \infty$ the youngest tip is always chosen for branching, leading to the comb-like tree. If $\alpha \to -\infty$, it is the oldest one which is always chosen, leading to a tree as symmetric as possible under the present rules.

We will consider first the case in which time increments are constant ($\Delta t = 1$, so that t in fact is equivalent to the number of species n; we assume that the root speciates at time $t = \Delta t = 1$) and later the case $\Delta t = 1/n$, which corresponds to the more realistic setting of a number of speciation events proportional to number of species in a constant time interval. Some of our results do not require to specify Δt as a function of time or of species number, so that the arguments could be applied to rather general cases.

4.4 Mean Depth Scaling

4.4.1 Numerical approach

Mean depth scaling for the age model defined in the previous section was first obtained numerically by averaging over 100 simulations for each value of α. Figure 4.2 is for $\Delta t = 1$ and Figure 4.3 is for $\Delta t = 1/n$. The main result of the figures is that

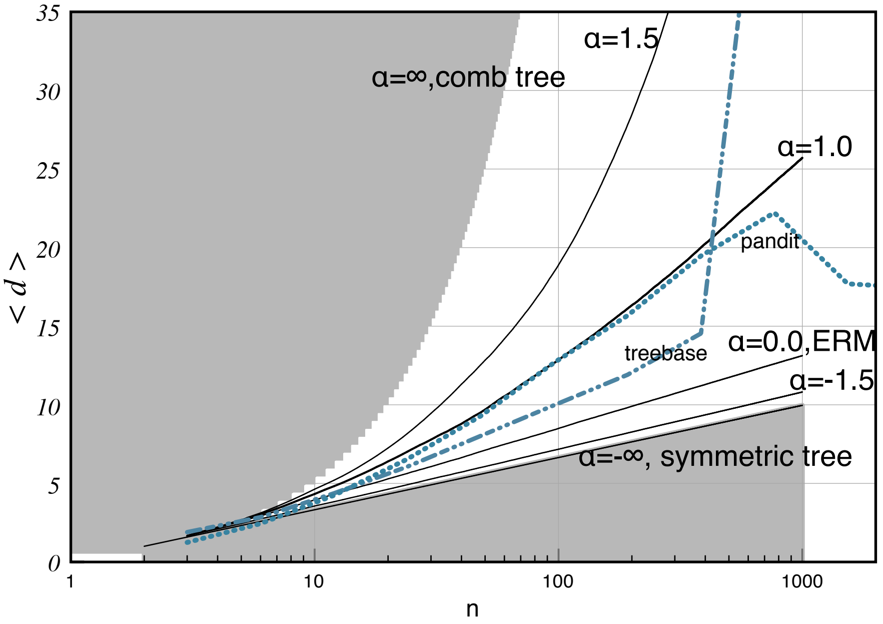

Figure 4.2. Figure shows the model's mean depth scaling ($\Delta t = 1$) for different α values (solid lines) in comparison to real phylogenetic trees (TREEBASE and PANDIT, blue dashed lines). Note that the scale of the plot is such that $\log n$ seems to be straight line.

the scaling for $\alpha = 1$ is $\langle d \rangle \sim log^2 n$ in both cases (with different prefactors) and are in good agreement with the scaling behavior observed in the databases. The $\log^2 n$ scaling of mean depth is mentioned in (Blum and François, 2006) in the context of the AB model, already stressing its aggreement to estimated data. When moving away from the $\alpha = 1$ case, the mean depth scaling rapidly goes in the direction of the behavior of the comb tree (mean depth scaling linear with n) for $\alpha > 1$, and to that of symmetric trees ($\langle d \rangle \sim \log n$) for $\alpha < 1$. Simulations for larger tree sizes are plotted in Figure 4.4 ($b = 0.0$ cases), supporting that the asymptotic behavior of the model for $\alpha = 1$ is $\langle d \rangle \sim \log^2 n$.

We checked the sensitivity of the inverse dependence $p_i \propto 1/\tau_i$ ($\alpha = 1$) by replacing it with $p_i \propto (\tau_i + b)^{-1}$ where b is a constant. The results can be seen in Figure 4.4. Only for large values of b deviations from the $\log^2$ scaling are evident (more important for the $\Delta t = 1/n$ case) so that the model is quite robust to perturbations of the $1/\tau_i$ branching probability dependence.

4.4.2 Analytical Approach

The simulation results above can be captured by some analytic arguments. The premise is that the average path from the tips to the root, which consists of $\langle d \rangle$ (mean depth) branches, is made of branches whose ages are approximated by their expected value at time t. In terms of $t(\langle d \rangle)$, the time at which the expected mean

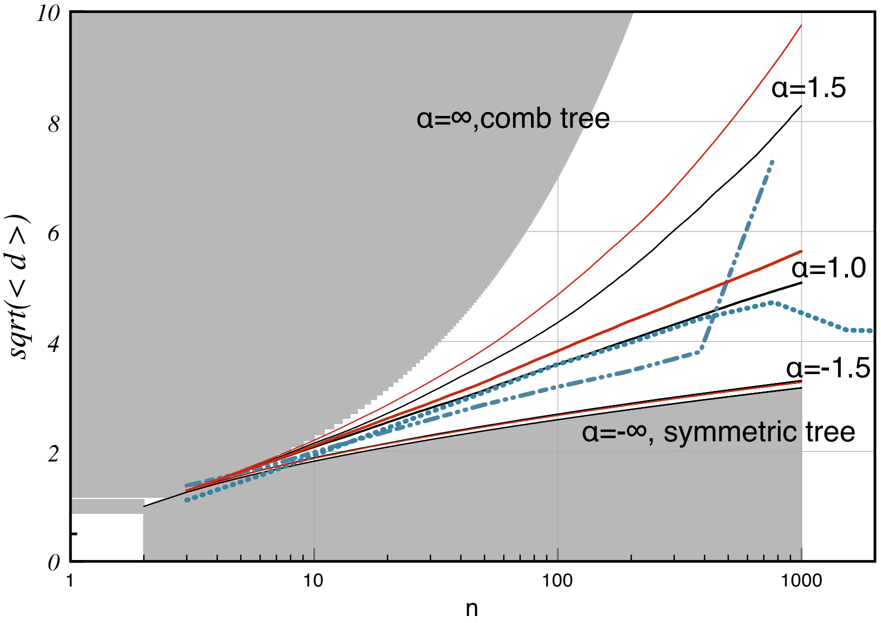

Figure 4.3. Comparison of mean depth scaling of the Aging model for $\Delta t = 1$ case (black lines) and for $\Delta t = 1/n$ (red lines). Note that the scale of the plot is such that $(\log n)^2$ seems to be straight line.

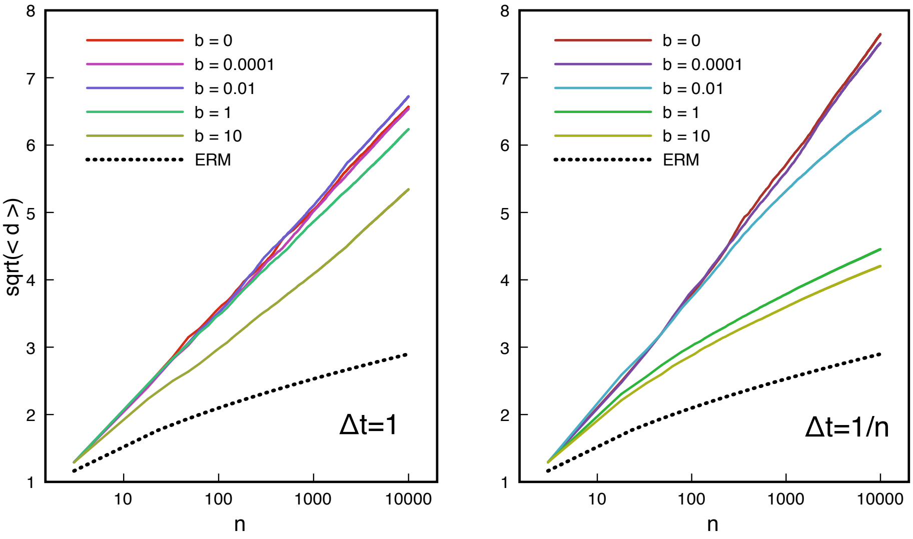

Figure 4.4. Deviation from $(\log n)^2$ mean depth scaling of the model for $\alpha = 1$ by changing the branching probability to proportional to $(\tau_i + b)^{-1}$.

depth $\langle d\rangle$ occurs, which is the function inverse of $\langle d\rangle = \langle d\rangle(t)$, this is written as:

$$t(\langle d\rangle) - t(\langle d\rangle - 1) = \langle\tau\rangle_{t(\langle d\rangle)}\ . \tag{4.4}$$

$\langle\tau\rangle_{t(\langle d\rangle)}$ is the expected age of the tip *chosen* at time $t(\langle d\rangle)$. When $\langle d\rangle \gg 1$ the difference can be approximated by a derivative:

$$\frac{dt}{d\langle d\rangle} = \langle\tau\rangle_{t(\langle d\rangle)} \tag{4.5}$$

We can now write the above expression in terms of n instead of t. For $\Delta t = 1$, $t = n$, so that

$$\frac{dn}{d\langle d\rangle} = \langle\tau\rangle_n \tag{4.6}$$

or equivalently

$$\frac{d\langle d\rangle}{dn} = [\langle\tau\rangle_n]^{-1}. \tag{4.7}$$

For $\Delta t = 1/n$, $t = \sum_{k=1}^{n} 1/k \approx \log n + \gamma + \ldots$ for big n (γ is the Euler constant), so that $dt/dn \approx 1/n$. Therefore:

$$\frac{dn}{d\langle d\rangle} = n\,\langle\tau\rangle_n \tag{4.8}$$

or equivalently

$$\frac{d\langle d\rangle}{dn} = [n\,\langle\tau\rangle_n]^{-1}. \tag{4.9}$$

We stress that the above expressions are valid for large $\langle d\rangle$ and n. More in general, for an arbitrary sequence of increments $\Delta t(n)$, $n = 1, 2, ...$, we have $t(n+1) - t(n) = \Delta t(n)$. For large n and t, we can approximate the difference by a derivative, leading to $dt/dn \approx \Delta t(n)$, so that (4.5) implies that the mean depth scaling will be given by

$$\frac{d\langle d\rangle}{dn} = \frac{\Delta t(n)}{\langle\tau\rangle_n}. \tag{4.10}$$

Now, we only need to estimate the behavior of $\langle\tau\rangle_n$, the average age of the chosen-tips as a function of n to determine the mean depth scaling of model. The expected number of tips having an age τ at time t is denoted by $m(\tau, t)$. This t should be considered as the time *just before* a branching occurs from n to $(n+1)$ tips.

Therefore, at time t

$$n = \sum_{\tau} m(\tau, t). \tag{4.11}$$

Considering that only those 2 tips which were just added Δt time ago have an age Δt, all other tips increase their age by Δt if they were not chosen for branching, and the chosen one has disappeared, then we can write recursive expressions for the expected number of tips of given age:

$$m(\tau,t) = \begin{cases} 2 & \text{if } \tau = \Delta t \\ \left(1 - \frac{c(t-\Delta t)}{\left(\tau-\Delta t\right)^{\alpha}}\right) m(\tau-\Delta t, t-\Delta t) & \text{if } \tau > \Delta t \end{cases} \tag{4.12}$$

The above expressions are for expected values, and in writing them we have assumed that the normalization factor $c(t)$ from (4.3) depends only on time, and not on the particular set of tip ages $\tau_1, ..., \tau_n$, i.e. we have used the mean value

$$[c(t)]^{-1} = \sum_{\tau} m(\tau,t)\tau^{-\alpha}, \tag{4.13}$$

which is a good approximation to (4.3) when there is a sufficiently large number of tips.

Expressions (4.12) and Eq.(4.13) can be iterated to obtain the tip age distribution for any n.

Now, we can write the average chosen-tip age in terms of m, (4.2) and (4.13):

$$\langle \tau \rangle_t = c(t) \sum_{\tau} m(\tau,t)\tau^{-\alpha+1} \ . \tag{4.14}$$

The numerical evaluation of this expression for the values of $m(\tau,t)$ obtained recursively before are shown in Figure 4.5 for our two elections of Δt.

We can understand the observed numerical behavior analytically, at least for the case $\Delta t = 1$. The main point is that, as revealed by extensive solutions of the recurrence equations, the dependence on τ and $t = n$ of $m(\tau,t)$ at large n converges to the scaling form:

$$m(\tau,n) = q(x = \tau/n) \tag{4.15}$$

Note that $x = \tau/n \in [1/n, 1)$, and for large n we can approximate it to be a continuous variable. Also, $q(x = 1/n) = m(1,n) = 2$ and $q(x = 1) = m(t,t) = 0$. Equations (4.11), (4.13), and (4.14) become, respectively:

$$1 = \int_{1/n}^{1} q(x)dx, \tag{4.16}$$

$$[c(n)]^{-1} = \int_{1/n}^{1} q(x)n^{-\alpha+1}x^{-\alpha}dx, \tag{4.17}$$

and

$$\langle \tau \rangle_n = c(n) \int_{1/n}^{1} q(x)x^{-\alpha+1}n^{-\alpha+2}dx \ . \tag{4.18}$$

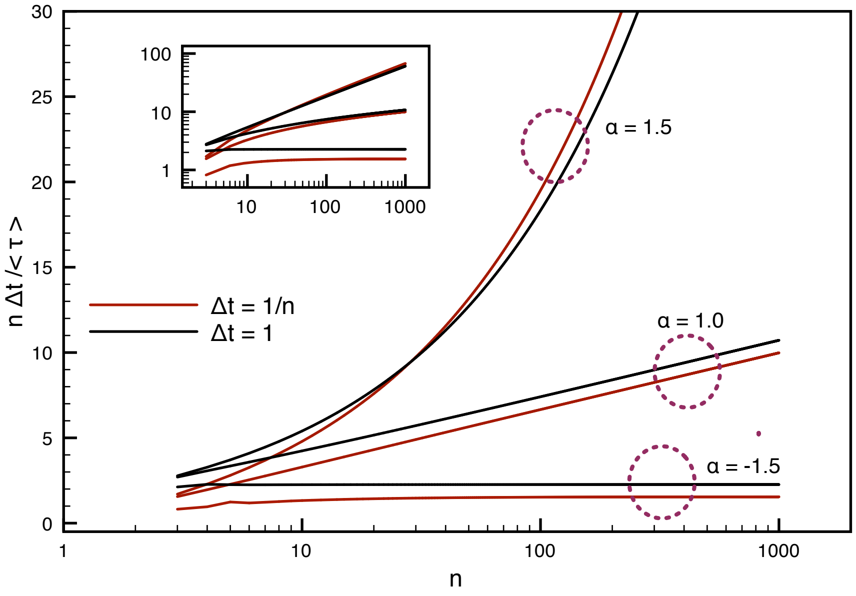

Figure 4.5. Scaling of the expected age of the chosen tip, as a function of the number of tips, obtained from the recursive Eq. 4.14.

The behaviour of these quantities as $n \to \infty$ can be elucidated by analysing the small-x behaviour of the integrand, and the fact that since $q(x = 1/n) = 2$, $q(x)$ remains finite as $x \to 0$. The result is

$$c(n) \sim \begin{cases} n^{\alpha-1} & \text{if } \alpha < 1 \\ (\log n)^{-1} & \text{if } \alpha = 1 \\ constant & \text{if } \alpha > 1 \end{cases} \tag{4.19}$$

and

$$\langle \tau \rangle_n \sim \begin{cases} n & \text{if } \alpha < 1 \\ \frac{n}{\log n} & \text{if } \alpha = 1 \\ n^{2-\alpha} & \text{if } 1 < \alpha < 2 \\ \log n & \text{if } \alpha = 2 \\ constant & \text{if } \alpha > 2 \end{cases} \tag{4.20}$$

This analysis is in good agreement with both simulations and recursive expressions as given in Figures 4.5 and 4.6.

Remembering Eq.(4.7) we conclude

$$\langle d \rangle \sim \begin{cases} \log n & \text{if } \alpha < 1 \\ (\log n)^2 & \text{if } \alpha = 1 \\ n^{\alpha-1} & \text{if } 1 < \alpha < 2 \\ \int dn/\log n & \text{if } \alpha = 2 \\ n & \text{if } \alpha > 2 \end{cases} \tag{4.21}$$

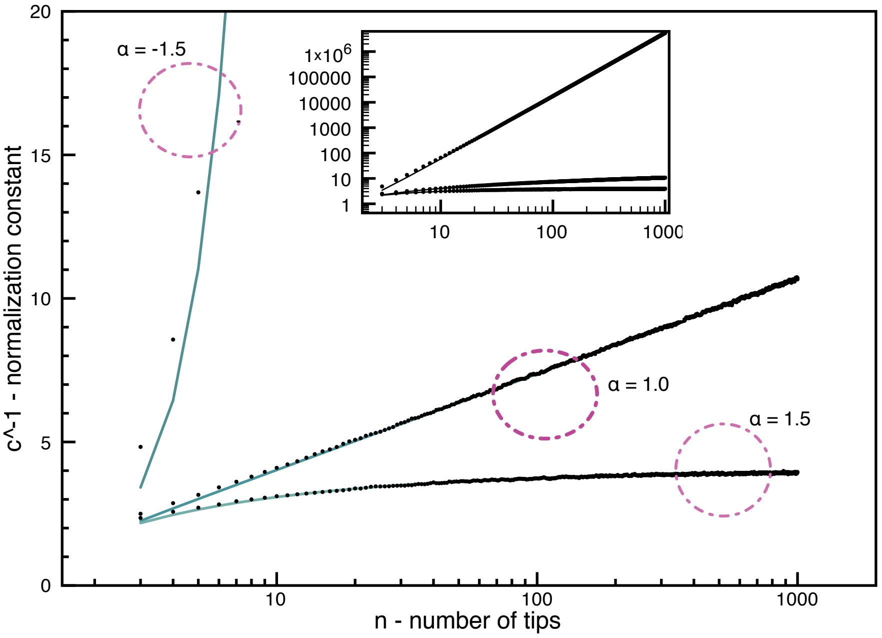

Figure 4.6. Comparison of the normalization constants ($\Delta t = 1$ case) from Eq.(4.13) (lines) and from direct simulation (points). Note that they agree with the approximate analytical results in Eq.(4.19).

We see that the aging model with $\alpha = 1$ has a scaling behavior consistent with the reconstructed trees in databases. Note that the known behavior of the symmetric tree and of the comb tree are correctly recovered in the $\alpha \to -\infty$ and $\alpha \to \infty$ limits, respectively.

4.5 Discussion

In macroevolutionary context, our model can be classified in a modeling scheme where species are units of dynamics (Erwin, 2000). In other words, species are taken as the smallest interacting bodies with some properties and due to those properties they are selected for speciation or extinction, like individuals in population genetics. For example, the geographical range is mentioned as an inheritable species' property in reference (Erwin, 2000). If an age property of species exists, then our model becomes a null-model in the sense that it explains the structure of the speciation (in the sense of balance scaling of the phylogenetic trees) with one parameter.

An additional interesting point of analysis and comparison of phylogenetic tree shape might be branching lengths distributions. TREEBASE and PANDIT do not contain this sort of data. However, even though our model is explicitly related to this quantity we have not intended here such comparison. The main reason of that branch length data are not as reliable as the topological structure of phylogenetic trees (Barraclough and Nee, 2001). This argument is supported in reference (Pigolotti et al., 2005) by summarizing the variety of behaviors of distributions found in the literature. We believe future studies will yield more reliable data for branching length, in other words lifetime of species, and therefore our model will be able to be judged further for its branching length distribution.

One of the main questions in phylogenetic studies is how often speciation occurs. This is analysed by plotting the number of lineages versus evolutionary time (Barraclough and Nee, 2001). In our model, this is actually an input for determining Δt; it is linear ($n = t$) for $\Delta t = 1$ and exponential ($n = e^t$ for big n) for $\Delta t = 1/n$. This ingredient of model can be further altered, e.g. $\Delta t = f(t, n)$, and estimated results can be obtained.

Funding

This work has been supported by the European Commission NEST Pathfinder initiative on Complexity through project EDEN (Contract 043251).

Acknowledgements

MT would like to thank Alejandro Herrada, Stephan Steigele, Joan Pons for their valuable discussions with him regarding biological evolution.

Bibliography

Agapow, P. M., Purvis, A., 2002. Power of eight tree shape statistics to detect non-random diversification: A comparison by simulation of two models of cladogenesis. Systematic Biology 51, 866–872.

Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97.

Aldous, D., 1996. Probability distributions on cladograms. In: Aldous, D., Pemantle, R. (Eds.), Random Discrete Structures. Springer, pp. 1–18.

Aldous, D., 2001a. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Statistical Science 16, 23–34.

Aldous, D., 2001b. Stochastic models and descriptive statistics for phylogenetic trees from Yule to today. Stat. Sci. 16, 23–34.

Aragon, J. L., Torres, M., Gil, D., Barrio, R. A., Maini, P. K., 2002. Turing patterns with pentagonal symmetry. Physical Review E 65, 1–9.

Balcan, D., Kabakçıoğlu, A., Mungan, M., Erzan, A., 2007. The information coded in the yeast response elements accounts for most of the topological properties of its transcriptional regulation network. PLoS ONE 2.

Banavar, J. R., Maritan, A., Rinaldo, A., 1999. Size and form in efficient transportation networks. Nature 399, 130–132.

Barraclough, T., Nee, S., 2001. Phylogenetics and speciation. TRENDS in Ecology & Evolution 16, 391–399.

Barrio, R. A., Aragon, J. L., C., V., Torres, M., I., J., Montero de Espinosa, F., 1997. Robust symmetric patterns in the faraday experiment. Physical Review E 56, 4222–4230.

Blum, M. G., François, O., 2006. Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. Systematic Biology 55, 685–691.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U., 2006. Complex networks: Structure and dynamics. Phys. Rep. 424, 175–308.

Brunet, E., Derrida, B., Simon, D., 2008. Universal tree structures in directed polymers and models of evolving populations. Physical Review E 78, 1–9.

Burlando, B., 1990. The fractal dimension of taxonomic systems. J. Theor. Biol. 146, 99–114.

Burlando, B., 1993. The fractal geometry of evolution. J. Theor. Biol. 163, 161–166.

Caldarelli, G., Caretta Cartozo, C., De Los Rios, P., Servedio, V. D. P., 2004. Widespread occurrence of the inverse square distribution in social sciences and taxonomy. Phys. Rev. E 69, 035101.

Capocci, A., Rao, F., Caldarelli, G., 2008. Taxonomy and clustering in collaborative systems: The case of the on-line encyclopedia Wikipedia. Europhys. Lett. 81, 28006.

Cavalli-Sforza, L. L., Edwards, A. W. F., 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21, 550–570.

Cracraft, J., Donoghue, M. J., 2004. Assembling the Tree of Life. Oxford University Press.

Darwin, C., 1859. On the Origin of Species. John Murray.

De Los Rios, P., 2001. Power law size distribution of supercritical random trees. Europhys. Lett. 56, 898–903.

Dietz, K., 2005. Darwinian fitness, evolutionary entropy and directionality theory. BioEssays 27, 1097–1101.

Eguíluz, V. M., Hernández-García, E., Piro, O., Klemm, K., 2003. Effective dimensions and percolation in hierarchically structured scale-free networks. Phys. Rev. E 68, 055102.

Erwin, D. H., 2000. Macroevolution is more than repeated rounds of microevolution. Evolution & Development 2, 78–84.

Ford, D. J., 2006. Probabilities on cladograms: introduction to the alpha model. Ph.D. thesis, Stanford University, available from `arXiv:math.PR/0511246`.

Frigaard, N.-U., Martinez, A., Mincer, T. J., Delong, E. F., 2006. Proteorhodopsin lateral gene transfer between marine planktonic bacteria and archaea. Nature 439, 847–850.

Garlaschelli, D., Caldarelli, G., Pietronero, L., 2003. Universal scaling relations in food webs. Nature 423, 165–168.

Gavrilets, S., 2003. Perspective: models of speciation: what have we learned in 40 years? International Journal of Organic Evolution 57, 2197–2215.

Goldenfeld, N., Woese, C., 2007. Biology's next revolution. Nature 445, 369.

Gregory, T., 2008. Understanding evolutionary trees. Evolution: Education and Outreach.

Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F., Arenas, A., 2003. Self-similar community structure in a network of human interactions. Phys. Rev. E 68, 065103.

Harding, E. F., 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. 3, 44–77.

Harris, T. E., 1963. The theory of branching processes. Springer-Verlag, Berlin, and Prentice-Hall, Inc., Englewood Cliffs, N.J., reprinted by Dover, NY, 1989 and 2002.

Hendry, A., 2009. Speciation. Nature 458, 162–164.

Hernández-García, E., Herrada, E. A., Rozenfeld, A. F., Tessone, C. J., Eguíluz, V. M., Duarte, C. M., Arnaud-Haond, S., Serrao, E., 2007. Evolutionary and ecological trees and networks. In: Descalzi, O., Rosso, O. A., Larrondo, H. A. (Eds.), Nonequilibrium Statistical Mechanics and Nonlinear Physics: XV Conference on Nonequilibrium Statistical Mechanics and Nonlinear Physics. Vol. 913 of AIP Conference Proceedings. AIP, pp. 78–83.

Herrada, E. A., Tessone, C. J., Klemm, K., Eguíluz, V. M., Hernández-García, E., Duarte, C. M., 2008. Universal scaling in the branching of the tree of life. PLoS ONE 3, e2757.

Higgs, P. G., Derrida, B., 1992. Genetic distance and species formation in evolving populations. Journal of Molecular Evolution 35, 454–465.

Huxley, J., 1942. Evolution: The Modern Synthesis. Allen & Unwin, London.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.

Kingman, J. F. C., 1982. On the genealogy of large populations. Journal of Applied Probability 19, 27–43.

Kingman, J. F. C., 2000. Origins of the coalescent: 1974-1982. Genetics 156, 1461–1463.

Klemm, K., Eguíluz, V. M., Miguel, M. S., 2005. Scaling in the structure of directory trees in a computer cluster. Phys. Rev. Lett. 95, 128701.

Klemm, K., Eguíluz, V. M., San Miguel, M., 2006. Analysis of attachment model for directory and file trees. Physica D 214, 149–155.

Kowald, A., Demetrius, A., 2004. Directionality theory: a computitonal study of an entropic principle in evolutionary process. Proc. R. Soc. B, 1–10.

Lloyd, S., 2009. A quantum of natural selection. Nature Physics 5, 164–166.

Matsen, F. A., 2006. A geometrical approach to tree shape statistics. System Biology 55, 652–661.

Mooers, A., Heard, S. B., 1997. Inferring evolutionary process from phylogenetic tree shape. The Quarterly Review of Biology 72, 31–54.

Patel, A., 2003. Mathematical physics and life. In: Misra, J. C., Goswami, A., Kumar, P. (Eds.), Mathematical Sciences Series Vol.4: Computing and Information Sciences: Recent Trends. Narosa Publishing House, India, pp. 271–294.

Pigolotti, S., Flammini, A., Marsili, M., Maritan, A., 2005. Species lifetime distribution for simple models of ecologies. Proc Natl Acad Sci U S A 102, 15747–15751.

Pigolotti, S., Lopez, A., Hernandez-Garcia, E., 2007. Species clustering in competitive lotka-volterra models. Physical Review Letters 98, 1–4.

Pinelis, I., 2003. Evolutionary models of phylogenetic trees. Proc. R. Soc. Lond. B 270, 1425–1431.

Powell, C. R., McKane, A. L., 2008. Predicting the species abundance distribution using a model food web. Journal of Theoretical Biology 255, 387–395.

Pusuluk, O., 2009. Quantum algorithms and genetic code. with private communication, 1–56.

Reznick, D., Ricklefs, E., 2009. Darwin's bridge between microevolution and macroevolution. Nature 457, 837–842.

Ricklefs, R., 2007. Estimating diversification rates from phylogenetic information. Trends in Ecology & Evolution 22, 601–610.

Rodriguez-Iturbe, I., Rinaldo, A., 1997. Fractal River Basins: Chance and Self-Organization. Cambridge University Press.

Roughgarden, J., Oishi, M., Akcay, E., 2006. Reproductive social behavior: Cooperative games to replace sexual selection. Science 311, 965–969.

Rozenfeld, A. F., Arnaud-Haond, S., Hernández-García, E., Eguíluz, V. M., Serrão, E. A., Duarte, C. M., 2008. Network analysis identifies weak and strong links in a metapopulation system. Proceedings of the National Academy of Sciences 105, 18824–18829.

Sackin, M., 1972. Good and bad phenograms. Syst. Zool. 21, 225–226.

Sella, G., Hirsh, A. E., 2005. The application of statistical physics to evolutionary biology. Proc Natl Acad Sci U S A 102, 9541–9546.

Spigel, M. R., 1970. Transformadas de Laplace. McGraw-Hill.

Stevens, P. S., 1974. Patterns in Nature. Little Brown & Co, Boston.

Stich, M., Manrubia, S. C., 2008. Topological properties of phylogenetic trees in evolutionary models. preprint.

Syvanen, M., 1985. Cross-species gene transfer; implications for a new theory of evolution. J. Theor. Biol. 112, 333–343.

Vetsigian, K., Woese, C., Goldenfeld, N., 2006. Collective evolution and the genetic code. PNAS 103, 10696–10701.

Volkenstein, M. V., 1994. Physical Approaches to Biological Evolution. Springer-Verlag.

West, G. B., Brown, J. H., Enquist, B. J., 1997. A general model for the origin of allometric scaling laws in biology. Science 276, 122–126.

Yule, G. U., 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Phil. Trans. R. Soc. Lond. B 213, 21–87.

Curriculum vitae

Murat Tuğrul

Personal Information

Date of birth:	April 1, 1983
Place of birth:	Tirebolu - Giresun - Turkey
Citizen of :	Turkey
E-mail:	mtugrul@ifisc.csic-uib.es, muratugrul@gmail.com
URL:	http://www.ifisc.uib-csic.es/∼mtugrul http://sites.google.com/site/muratugrul/

Education

- (since February 2008)   Graduate Student in Physics.
  Institute for Cross-Disciplinary Physics and Complex Systems, Universitat de les Illes Balears, Palma de Mallorca.
  Advisors: Emilio Hernández García. & Víctor M. Eguíluz

- (December 2007)   M.Sc. in Computational Sciences & Engineerings.
  Koç University, Istanbul, Turkey
  Advisor: Alkan Kabakçıoğlu.
  Thesis Title: "*The Structure and Dynamics of Gene Regulation Networks*", arXiv:0802.1989 [q-bio.MN]

- (June 2005)   B.Sc in Physics (major) and Philosophy (minor).
  Middle East Technical University, Ankara, Turkey.

Publications

1. **Murat Tuğrul**, Stephanie Keller-Schmidt, Víctor M. Eguíluz, Emilio Hernández-García and Konstantin Klemm. "*Might an Age Dependent Branching Model Help Us to Understand Macroevolution?*", work in progress.

2. **Murat Tuğrul** & Alkan Kabakçıoğlu. "*Anomalies in the transcriptional regulatory network of Saccharomyces Cerevisiae*", submitted for publication, arXiv:0810.3877 [q-bio.MN].

3. Emilio Hernández-García, **Murat Tuğrul**, E. Alejendro Herrada, Víctor M. Eguíluz and Konstantin Klemm. "*Simple Models for Phylogenetic Trees*", to appear in IJBC, arXiv:0810.3877 [q-bio.QM].

4. **Murat Tuğrul** & Alkan Kabakçıoğlu. "*Robustness of Transcriptional Regulation in Yeast-like Model Boolean Networks*", to appear in IJBC, arXiv:0902.4147 [q-bio.MN].

Presentations

[O] = Oral presentation, [I] = Invited, [P] = Poster presentation

1. "*Structure and Dynamics of Yeast Gene Regulation*", Ecological Diversity and Evolutionary Networks project workshop, Leipzig, Germany, (2008). [O]

2. "*Boolean Dynamics of Gene Regulation Network of Saccharomyces Cerevisiae (yeast)*", Istanbul Statistical Physics Days, Istanbul, Turkey, (2008). [O]

3. "*Boolean Dynamics of Gene Regulation Network of Saccharomyces Cerevisiae (yeast)*", International Conference of Modelling and Computation on Complex Networks and Related Topics, Pamplona, Spain, (2008). [P]

4. "*Structural & Dynamical Aspects of Transcriptional Regulation in YEAST Genetic Network*", IFISC seminar, Palma de Mallorca, Spain, (2007). [O-I]

Teaching Experience

1. (2005-2007) Teaching Asistant in Koç University; PHYS 102 General Physics-2, SCIE 109 Physics of Everyday Life, MATH 204 Calculus-2, MATH 101 Discrete Mathematics.

2. (2006-2007) Priviate tutoring; mathematics, physics in high school and university first year level.