Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Hua Fang April 6, 2021

The Structural Asymmetry and Scaling of Phylogenetic Trees

by

Hua Fang

Dr. Stefan Boettcher
Adviser

Department of Physics

Dr. Stefan Boettcher

Adviser

Dr. Effrosyni Seitaridou

Committee Member

Dr. Daniel Weissman

Committee Member

2021

The Structural Asymmetry and Scaling of Phylogenetic Trees

By

Hua Fang

Dr. Stefan Boettcher

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Physics

2021

Abstract

The Structural Asymmetry and Scaling of Phylogenetic Trees

By Hua Fang

Understanding the patterns and processes of evolution is very challenging. Because they carry inherent information about evolution, phylogenetic trees become a prevalent tool in evolutionary biology. In this work, we studied large-scale phylogenetic trees in terms of tree asymmetry and distribution of branch length. We used two indicators: the ratio of the size of smaller child clade to the size of parent clade and the ultrametric distances between two consecutive nodes. By comparing with a random null model and a critical model, we found that both the topology and timing of trees are scale-invariant and could be described by a power law distribution. This scale-invariance suggests that similar forces drive the evolution over a large range of scales. However, the observed patterns of several trees are better described by the null tree from the Markov process. Future works could attempt to explain this deviation.

The Structural Asymmetry and Scaling of Phylogenetic Trees

By

Hua Fang

Dr. Stefan Boettcher

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Department of Physics

2021

Acknowledgements

I would like to thank Dr. Stefan Boettcher for providing me his wisdom, pushing me to learn more, and guiding me through this process. I would also like to thank Dr. Effrosyni Seitaridou and Dr. Daniel Weissman for taking their time to serve on my committee. My sincere appreciation and gratitude go to my honor committee members for their kind support and encouragements throughout the project, without which I would not have made it through.

Table of Contents

List of Tables and Figures

1.Introduction

One important aim of evolutionary studies is to explore the mechanism that leads to the current evolution's results. While there are a bunch of different theories, it is hard to get empirical or experimental evidences to resolve the chaos because the evolutionary process is very difficult to observe in real time. Phylogenetic trees, the tree diagram describing relationships among species based on their physical and genetic properties, are central to approach this issue. Because of the development of genetic sequencing, statistical tools, and databases over the recent decade, the quantity, size, and detail of available phylogenetic trees are surging, which in turn facilitate more sophisticated models. Analyzing the phylogenetic structure helps to draw inference on underlying principles such as detection of adaptive radiation and mass extinction, speciation and extinction rates, and ecological causes that lead to diversity of life (Mooers & Heard, 1997). Phylogenetic trees might also be interesting for physicists to study. For example, studies have investigated the percolation of fractal trees (Vandewalle & Ausloos, 1997) and phase transitions in tree reconstruction in terms of a random cluster model (Mossel & Steel, 2004) and by making an analogy to simulated annealing (Strobl & Barker, 2016).

In this paper, we analyzed the topology of phylogenetic trees in hope that it may shed light on inferring the macroevolutionary process. A key observation is that the structure of phylogenetic trees is more asymmetric than expected. It leads to questions such as whether this asymmetry is caused by pure chance or by some mechanisms, and by which mechanisms if any. To answer this question, various studies have attempted to develop metrics for measuring asymmetry (Blum & Francois, 2005; Colijn & Plazzotta, 2018; Liu et al., 2020), compare random models with phylogenetic trees (Matsen, 2006),

determine factors contributing to the tree shape (case study of historical time periods: Bernardi et al., 2016; ecological saturation and landscape: Gascuel et al., 2015; tree size and depth: Purvis & Agapow, 2002), and build evolutionary models to reproduce tree asymmetry (neutral biodiversity theory: Davis et al., 2011; niche construction: Xue et al., 2020). There are two aspects in studying tree shapes: topological balance and distribution of branch length (Mooers & Heard, 1997). Many studies have supported a scale-invariant relation in the topologies and metrics of trees (Herreda et al., 2008; Hernandez-Garcia et al., 2010). Theoretical studies and critical models of evolution have discussed the long-term memory effect as well (Bak & Sneppen, 1993; Boettcher & Paczuski, 1996). Here, we aimed to study the scale-invariant behavior of tree topology and branch length of 11 reconstructed phylogenies. We would compare the 11 trees to a tree generated by the Markov process and a tree generated by the critical model of Boettcher and Paczuski (1996).

1.1 Definitions about phylogenetic trees

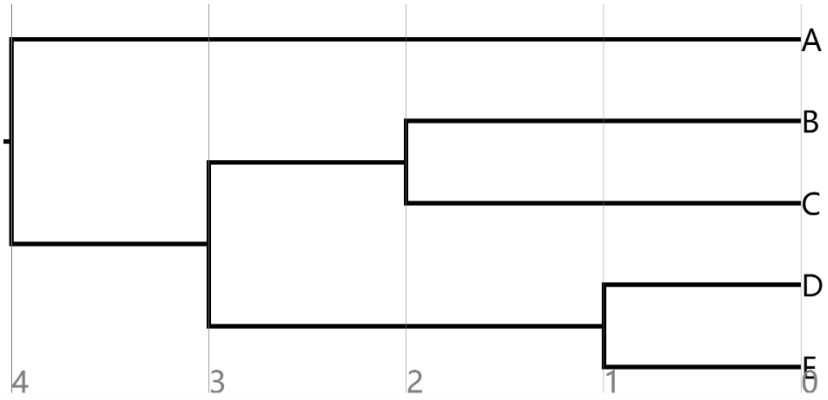


Figure 1. A time-scaled phylogenetic tree. The root is the small tick on the left and the five leaves are on the right. The horizontal axis is the age of each node. The branch length is the time elapsed since speciation.

Phylogenetic trees are usually represented by tree-like diagrams consisting of nodes and branches. The root, internal nodes, leaves, branch lengths, and clade are terms to describe a tree. There are also different types of trees. The type of trees used in this study is rooted, binary, and labeled, with branch length scaled by time. Figure 1 is an example of this type and it shows the phylogenetic relation between the root and five currently living organisms. First, a rooted tree is a tree with a root, which is represented by a single branch at the left. The root is the common ancestor, or the parent, of all the nodes in this tree. Its descendants are at the right. The tree is directional that from left to right, the time evolves from ancient to current. There are two kinds of nodes: internal and terminal. Terminal nodes are nodes without descendants, such as A, B, C, D, and E, and are often called leaf. Nodes with descendants are called internal nodes. They are the ancestors of current living species. Second, at each node, there are exactly two immediate descendants, which makes it a binary tree. Third, in a labelled tree, usually only the leaves are labelled with the name of the species. Fourth, the branches, shown by the horizontal lines, indicate the proximity of a node to its ancestor. Branch lengths are proportional to the time elapsed since the nodes branched off from the ancestor. A longer branch means a node is more distant in time from its ancestor. The distance from one node to the root can be obtained by summing all the branches from the node to the root. A clade is defined as a subtree rooted at an internal node. The size of a clade is the number of all the nodes in the subtree.

1.2 The critical model by Boettcher and Paczuski

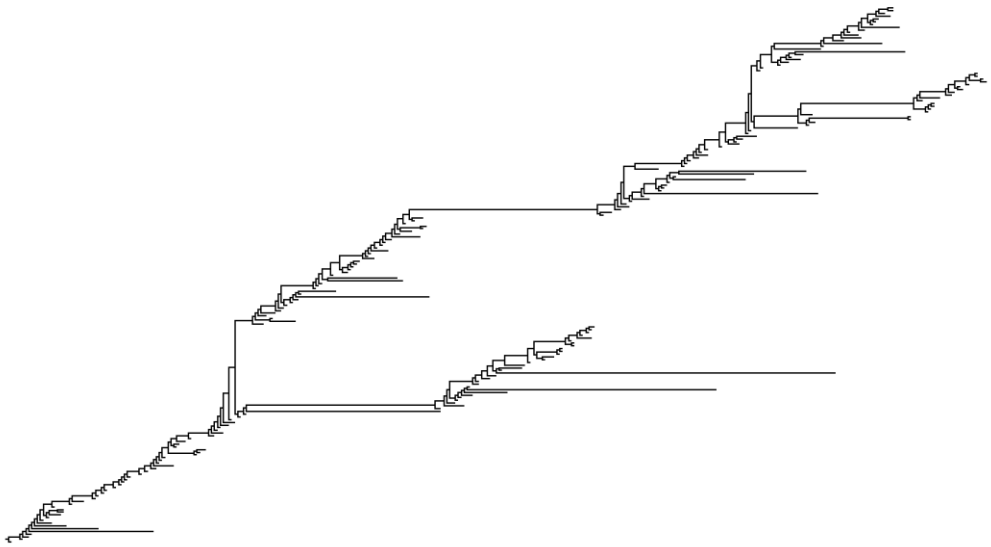


Figure 2. A tree generated by the model of Boettcher and Paczuski.

Based upon the self-organized critical model by Bak and Sneppen (1993), Boettcher and Paczuski (1996) developed a multi-trait evolution model where the species with the least fit mutate and may cause a chain reaction of coevolution. This model will be referred as the simulation model. Figure 2 is a visualization of a tree generated by this model, which is highly asymmetric.

The model consists of a lattice of species. Each species is assigned with multiple values. Those values represent their traits and how those traits help them to survive. A small value means the species is less fit to the environment and susceptible to mutation. In their multi-trait evolution model, each time, an individual species with the lowest fitness mutates. Taking inter-species interaction into account, this mutation will change one of the traits of two neighboring species with the potential of undermining their fitness. It reflects the idea of a feedback system between organisms and the environment.

Organisms are changed by the environment, but they can also modify the environment through mechanisms like food chains.

This system evolves by itself to critical states where intermittent avalanches of mutation happen between long periods of quiescence. This system supports punctuated equilibrium against phyletic gradualism. Phyletic gradualism argues that species evolve and diversify slowly and gradually with time. In punctuated equilibrium, long periods of equilibrium are interrupted by intermittent large-scale events that are called avalanches. Fossil records show that the traits of a particular species remain the same for long periods of time, but may experience drastic changes in relatively short periods.

In this model, activities during each avalanche are represented by a tree structure. The parent of a mutated species is the one that introduces a change on the specie's trait and makes the species the lowest fit.

1.3 The Markov model

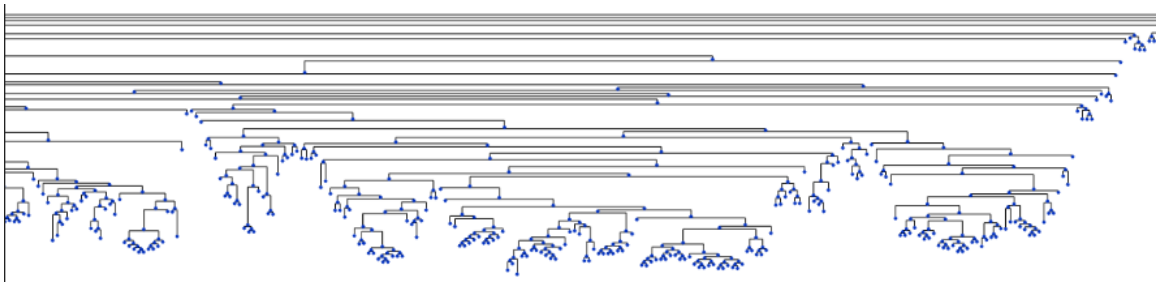


Figure 3. Visualization of a section of a large Markov tree used in this study. The blue dots correspond to leaves.

Our null model is generated by the Markov process and it is called as the Markov model. A Markov process is a stochastic process that the probability of an event only depends on its current state. Therefore, the future is independent of the past. An example

is the Brownian motion. The branch lengths are assigned according to a Poisson distribution. This process is a memoryless process, so the distribution of branch lengths should exhibit an exponential tail rather than a long-tailed power law. In addition, a visualization of the Markov tree in figure 3 shows that it is relatively symmetric.

1.4 Asymmetric tree structure

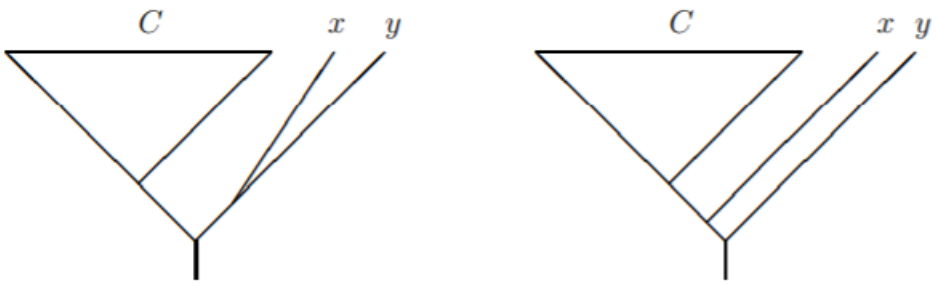


Figure 4. (Figure 1 in Jones, 2011) The two possible topologies to arrange a clade (C) with n tips and two tips (x, y) that do not belong to C. Left: a more balanced structure. Right: a less balanced structure. The ratio of probability of left to right is n under the ERM model and 1 under the PDA model.

The shape of phylogenetic trees refers to the connectivity of nodes without taking the branch lengths into account. The shape of phylogenetic trees carries information about the evolutionary process. An interesting phenomenon that draws a lot of attention is the asymmetry of tree structures. Figure 4 is a graph by Jones (2011). In a rooted tree with n+2 tips, there is a clade (C) containing n tips and two tips (x, y) sitting outside C. Figure 4 shows the two possible topologies of such a tree, where the left topology is more balanced than the right one. The simplest model of branching process with continuous time was proposed by Yule in 1925. The Yule model is an equal-rate Markov (ERM) branching process, where each branch has an equal probability of splitting. Under ERM,

the ratio of the probability of the left to the right topology is n. Since ERM usually gives more balanced trees, it is often used as a null hypothesis in current studies. Another early model is the proportional-to-distinguishable-arrangements model (PDA). It makes all tree topologies equally likely, resulting in a less balanced tree. Both models have received interpretations and justifications from evolutionary perspectives.

The two models were integrated into a more generalized model of cladogram distributions developed by Aldous (2001). In a one-parameter beta-splitting family with $-2 < \beta \leq \infty$, the ERM and the PDA model are special cases corresponding to $\beta = 0$ and $\beta = -1.5$, respectively. Higher β corresponds to greater balance. Aldous and early studies agreed that reconstructed phylogenies are more balanced than trees predicted by PDA, and less balanced than trees predicted by the ERM model. This tendency is independent of "methodological details of the trees' estimation" (Heard 1996), meaning that the different ways to reconstruct trees from evolutionary data cannot cause or remove this tendency.

Over the recent years, there are numerous models attempting to establish links between tree asymmetry and various factors such as geographical factors, access to resources, and rate of diversification (Morlon, 2014). Many different metrics have also been developed to characterize and compare tree structures. Given the increasing size of trees, a particular parameter can hardly give a comprehensive description. Although there is a sizeable literature of statistics measuring tree balance and distribution by using the Markov model as a null hypothesis, it is unsatisfactory for the overall properties of a collection of trees with different size (Aldous 2011). In this paper, we studied a collection of large trees by applying a less arbitrary but powerful method proposed by Aldous

(2001), who was unable to analyze large phylogenetic trees then. This method uses the ratio of size of smaller child clade to size of parent clade as an indicator.

A clade can be understood as a subtree rooted at an internal node. The basic idea of this method is to count the number of nodes for a subtree rooted at an internal node i (the count includes node i itself). For a binary tree, each internal node specifies the split of a parent clade into two child clades. Figure 5 is a graph by Xue et al. (2020) to demonstrate the symmetry of binary trees. On the left of figure 5 is a balanced tree, where two child clades of any parent clade are equal in size. The child clade size in a balanced tree determines the upper limit of a smaller child clade size. On the right is a maximally unbalanced tree, which gives the lower limit of the smaller child clade size as the smaller of the two child clades always has a size of one. Let the size of the parent clade (number of nodes within the subtree rooted at the parent node i) be $P(i)$, and let the size of the smaller child clade be $S(i)$, then $1 \leq S(i) \leq \frac{P(i)-1}{2}$. There is a -1 because the parent node i is counted, but it does not belong to any child clade.

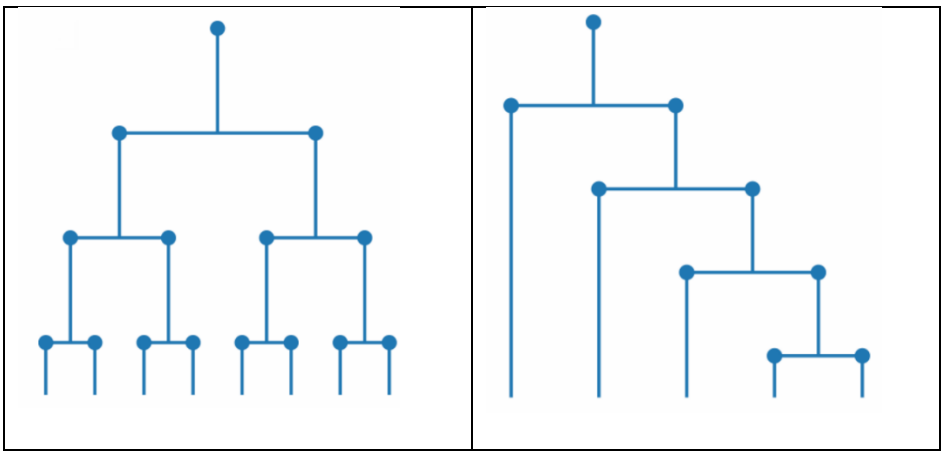


Figure 5. (Fig. 1 by Xue et al., 2020) Left: 1. A balanced tree. Each parent splits into two child clades with equal sizes. Right: 2. A maximally unbalanced tree. Each smaller child clade has a size of one regardless of the size of the parent.

A scatter plot of smaller child clade size over parent clade size is graphed for each internal node. A flatter slope indicates a higher degree of asymmetry. The slope can be compared with graphs generated by the two models to infer the underlying evolutionary process. As shown in figure 2 and figure 3, the Markov tree is more balanced than the simulation tree. We predicted that in the scatter plot, a Markov tree would have a steeper slope than the simulation tree. The Markov tree should also be steeper than most natural trees because it is shown to be more balanced than natural trees by various studies.

1.5 Ultrametricity

A metric is a function that defines the distance between any two points in a given set of objects *M*. A metric d on *M* maps the set to a real number $d(x, y)$ such that for any $x, y, z \in M$:

$$d(x, y) > 0 \quad \text{for } x \neq y,$$

$$d(x, y) = 0 \quad \text{for } x = y,$$

$$d(x, y) = d(y, x),$$

$$d(x, y) \leq d(x, z) + d(y, z),$$

where the last condition is called the triangle inequality. An ultrametric satisfies the ultrametric inequality, which can be seen as a strengthened triangle inequality:

$$d(x, y) \leq \max\{d(x, z), d(y, z)\}.$$

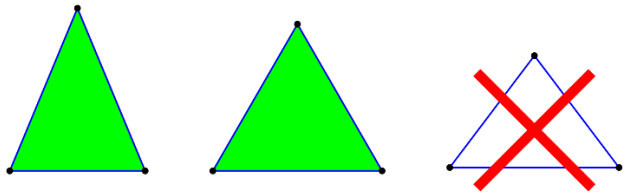


Figure 6. (“Ultrametric space”, n.d., sec. “Properties”) A graphical illustration of ultrametric inequality. The ultrametric inequality is satisfied by an isosceles triangle with a shorter base (the left) and an equilateral triangle (the middle), but not an isosceles triangle with a longer base (the right).

Figure 6 shows graphical examples of the ultrametric relations. The ultrametric inequality is satisfied by triangles in the left and in the middle while it is not satisfied by the triangle in the right. For three arbitrary points, there are three distances between them. The ultrametric inequality demonstrates that one of the three distances must be smaller or equal to the remaining two distances. For example, when at least two distances are equal, it is an isosceles triangle with a shorter unequal side (figure 6, left). When all three distances are equal, it is an equilateral triangle (figure 6, middle). In the right of figure 6, the isosceles triangle with a longer base than its sides does not satisfy the inequality. Distance satisfying the ultrametric inequality is ultrametric distance.

There is a wide range of applications of ultrametricity. Rammal et al. (1986) gives a comprehensive review of ultrametricity from a physicist’s perspective, which discusses applications in taxonomy, spin glasses, random walk, simulated annealing, neural networks, and protein freezing and folding. Since phylogenies are issues of biological taxonomy with the use of hierarchical trees, it is natural to consider ultrametricity for phylogenetic analysis.

For a rooted tree with a time axis, all the nodes that live at the same time moment satisfy the ultrametric inequality because the time elapsed (i.e. the distance) from the root to those nodes is the same (Rammal et al., 1986). The distance between two nodes is determined by their lowest common ancestor (LCA), which is their first common ancestor back in time. For any three nodes (A, B, C), two nodes, say A and B, that share a younger LCA (M) can be seen as a cluster. The distance between M and the cluster is the shorter base (d (A, M) = d (B, M)). If the LCA of C and the cluster is N, then the distance from C to N or the cluster to N forms the isosceles sides (d (A, N) = d (B, N) = d (C, N)). Finally, as M is closer to A and B, d (A, N) = d (B, N) = d (C, N) > d (A, M) = d (B, M). Thus, for two nodes that have the same distance to root, the distance is ultrametric. If M and N is the same node, then the three nodes A, B, and C have the same LCA and their distances are equal.

Boettcher and Paczuski (1996) developed a multitrait evolution model where the species with the least fit mutate and may cause a chain reaction of coevolution. Because this simulation used equal time steps, the distance between any two nodes in this tree at a given time is ultrametric. They demonstrated that the ultrametric tree structure of avalanches, which measures the ultrametric distance between successive events in the chain reaction, follows a power law distribution at large times. It revealed the history-dependent property of avalanches, meaning that the events further away from the root of tree correlate with the events happened a long time ago. In contrary, for the null tree from the memoryless Markov process, its ultrametric distribution should be Poisson instead of a power law.

Other studies have directly introduced a long-term memory into the branching process by requiring the branching probabilities to be a power law function of branching age (Keller-Schmidt et al., 2015). Through this way, the same power-law scaling could be observed. However, it could not provide an evolutionary interpretation of the observed pattern because it is forcing the simulation to produce expected or known results.

This study applies the idea in Boettcher and Paczuski to reconstructed phylogenetic trees. Generally, in natural trees, it is rare that internal nodes are at the same time moment because branching rarely happens at the exact same moment. Thus, the distances from internal nodes to their LCAs are not exactly ultrametric. In this study, given that the phylogenetic trees are relatively large, the time difference between consecutive nodes are relatively small. Thus, the distance between consecutive branching points is approximately ultrametric. In conclusion, the ultrametric distribution should be exponential for the Markov tree and power-law for the simulation tree. For the phylogenetic trees, if their evolutionary events have memories about the past, their ultrametric distances should also be described by power laws.

2. Methods

2.1 Newick Tree Format

The Newick format is a way of representing trees in computer-readable forms by using commas and nested parentheses. It resembles the high-level programming language Lisp. The Newick format is a minimal definition for phylogenetic trees, where only commas, parentheses, and a semicolon are needed to topographically represent a tree

structure. Besides, the name of nodes and edge lengths can be added into the basic structure by using characters, numbers, and colons. For example, the tree in figure 1 in Newick format is "*(((E:1.0, D:1.0):2.0, (C:2.0, B:2.0):1.0):1.0, A:4.0):0.0;*". A colon is used to separate names and lengths. The characters preceding the colon are the names of nodes. The numbers following the colon are the edge lengths between a node and its parent. If a node is not labelled, there is no character before the colon. The semicolon indicates the root of tree. The root has length zero because it does not have a parent. Nodes inside a matched pair of brackets and separated by commas are siblings, the immediate descendants of the same parent node. Following the right bracket is the parent of nodes inside the bracket. For example, the unnamed internal node with length two is the parent of D and E.

2.2 Data

Two types of trees are analyzed here: phylogenetic trees reconstructed by biology tools and artificial trees simulated by the model of Boettcher and Paczuski. All the trees are binary and rooted. 11 reconstructed trees are downloaded from the online database TreeBASE (https://treebase.org), a repository of phylogenetic information in peer-reviewed publications. The 11 trees are determined after looking for trees which are larger than 1000 taxa in TreeBASE and readable by our program. Table 1 in the appendix presents a description of the downloaded trees.

Trees are downloaded as NEXUS format and converted to the Newick format. NEXUS is a file that consists of "separate blocks, each containing one particular kind of information", such as information about "taxa, morphological and molecular characters, distances, genetic codes, assumptions, sets, trees, etc." (Maddison, et al., 1997).

Basically, the Newick format can be seen as a block in the Nexus format that contains phylogenetic information.

2.3 Implementation

A program is developed in Spyder Python 3.7 that performs basic operations such as parsing and traversing, as well as statistical analysis such as calculating the ultrametric distances. Numpy, matplotlib, scipy.stats, and math are imported. For convenient parsing of trees, trees are converted to Newick format via the Python package Bio.Phylo. The Newick file is parsed iteratively. Each species is stored as a type Node with pointers to its parent and two immediate descendants. The root node is returned after parsing. For multiple trees in one file, a list of roots is returned. By supplying the root into the Preorder traversing function, all the nodes are returned in a list.

2.4 Size of parent clade and child clade

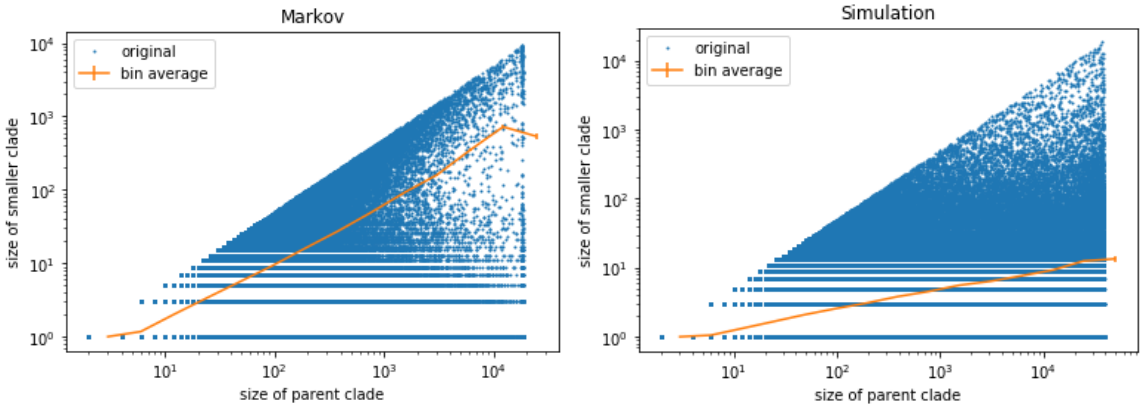
A method that implements stack is used to get the size of a subtree rooted at any input node. Then, while traversing over the entire tree, the size of the subtree at each internal node and the size of its smaller child subtrees are obtained. Two arrays storing the sizes are used to plot the original scatter plot. The vertical and horizontal axes are scaled to log of base 10. In addition, data points of parent clade size are grouped via log-binning of base 2. Log2 is chosen because it is appropriate for the data value range and the number of data points. The number of bins depends on the maximum value of points and is approximately ten for trees studied. The mean and standard deviation of each bin are calculated. The estimated standard error (the standard deviation divided by the square root of frequency) is plotted on the original scatter plot.

2.5 Ultrametric distance

To calculate the ultrametric distances, the lowest common ancestor (LCA) of any two nodes is found. Then, the distance from root to every node is calculated and the nodes are sorted in terms of increasing distance. For each pair of nodes that are consecutive in time, two distances from the two nodes to their LCA are recorded in an array of ultrametric distances. Log-binning of base 2 is applied to group the data into ten bins. Lastly, a probability density of ultrametric distances is plotted in base 10 logarithmic scale.

3. Results and Discussion

3.1 Tree topology



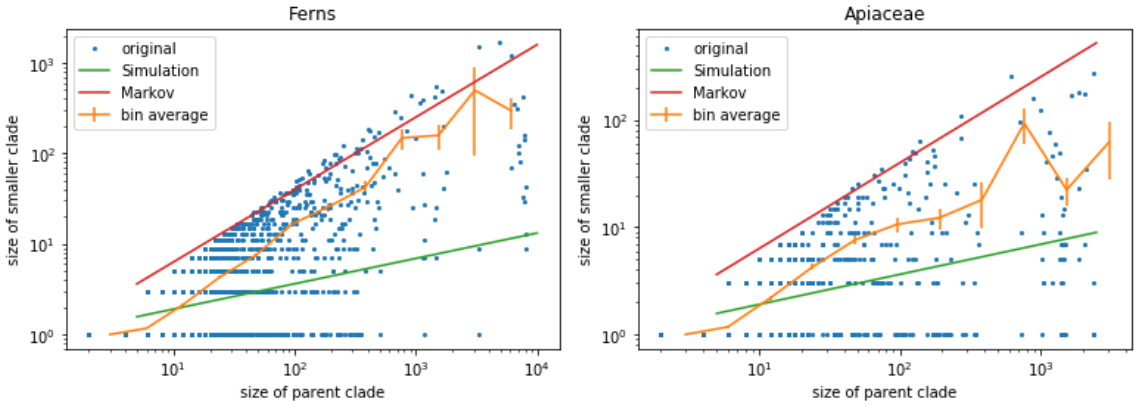


Figure 7. Log-log plots of the size of smaller child clade versus the size of parent clade. (1) Top left: a null tree generated by the Markov process. (2) Top right: a simulated tree generated by the program of Boettcher and Paczuski. (3) Bottom left: the tree of ferns is relatively balanced because its bin average has a relatively large slope. (4) Bottom right: a relatively unbalanced tree of apiaceae because a relatively small slope of bin averages.

Figure 7 shows the asymmetry information of two simulated trees and two natural trees. The size of smaller child clade was plotted over the size of parent clade with log-log scale. The blue dots represent the original scatter plot of correspondence between the horizontal and the vertical axis. The blue dots were grouped by log-2 bins and the average of each bin was calculated and connected by the orange line. Centered at each average point, vertical error bars were plotted, but the standard error was too small to be seen on the graph because of the large number of data points.

The slope indicates the level of symmetry. A steeper slope corresponds to a higher level of symmetry while a flatter slope corresponds to a lower level of symmetry. In figure 7, all data points are in the region between the diagonal and y = 1. In a log-log plot, the slope of a line is the exponent of a power law. As discussed before, the maximum

value of a smaller clade is $S(i) = \frac{P(i)-1}{2}$, where P is the size of parent clade rooted at node i. Therefore, a balanced tree should be represented by approximately y = 0.5x on this graph, which means that the upper limit is the diagonal with a slope of one. The lower limit is a flat line y = 1. It is given by a maximally unbalanced tree where the size of the smaller child clade is always 1 regardless of the parent clade size.

The tree in figure 7.1 is the null tree (referred as the Markov tree) generated by the Markov process. The tree in figure 7.2 is a very unbalanced tree (referred as the simulation tree) generated by the simulation algorithm in Boettcher and Paczuski (1996). We found the equations that fit the two lines of average bins are: $y_o = 4x_o^{0.8}$ for the Markov tree, and $y_s = 1.7x_s^{0.28}$ for the simulation tree. The slope of bin averages signifies how much asymmetry is present. A slope that is steeper and closer to the diagonal is more balanced. It is obvious that the slope of the Markov tree is steeper than that of the simulation tree, so the Markov tree is more balanced. In addition, we validated our Markov tree by finding that its slope is parallel to the slope of Aldous's Markov tree (figure 12 in appendix).

First, we found that all natural trees in our study fall in the region between the Markov tree and the simulated tree. Graphs for all the natural trees are attached in the appendix as figure 13. Figure 7.3 and 7.4 are two natural trees of ferns and apiaceae, respectively. The slopes of the Markov and simulated trees are represented by the red and green lines, respectively. In general, most natural trees in this study have more proximity to the Markov slope than the simulated slope. Fig. 4 and fig. 5 in Aldous (2001) provide a closer look (figure 12 in Appendix). Roughly, all of our natural trees lie in the region between $\beta = 0$ and $\beta = -1$. The results show that the evolutionary conditions of the

simulated tree are too extreme for natural trees. For example, the ratio of speciation to extinction rates is too small or too large. Second, the level of asymmetry varies between trees. The fern tree is one of the most balanced trees used in our study. Its slope is roughly parallel or even greater than the Markov slope. The apiaceae tree in figure 7.4 is an example of a relatively unbalanced tree. The region of smaller x-values resembles the Markov slope more while the region of intermediate values resembles the simulated tree. Besides, comparing to the computer-generated trees, connecting bin averages in natural trees does not give a perfect straight line. It is caused by factors such as different speciation-extinction rates between species or genera in natural trees (Aldous 2011).

3.2 Ultrametric distance

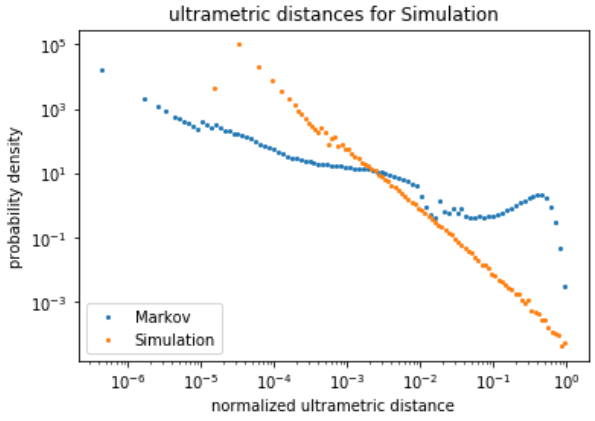


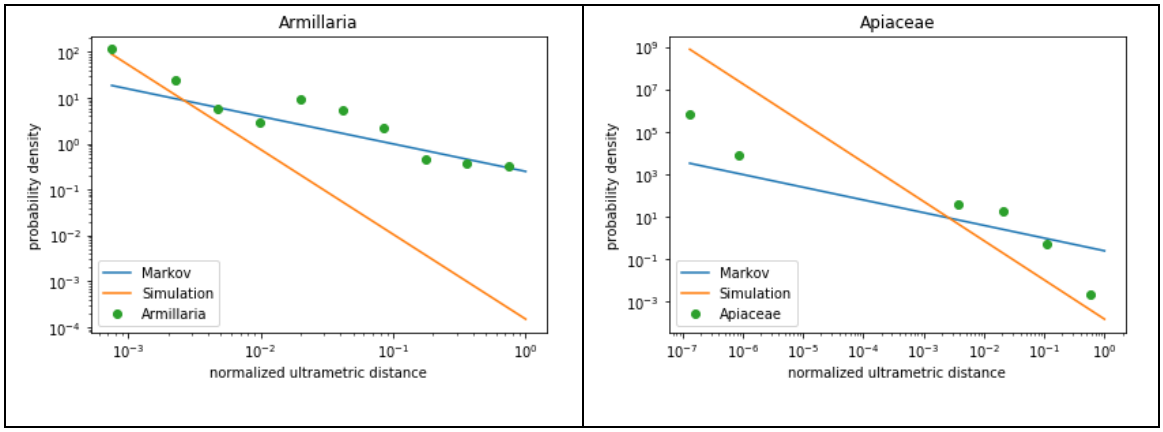
Figure 8. Log-log plots of distribution of normalized ultrametric distances for computer-generated trees with 100 bins. Blue dots: a null tree generated by the Markov process. Orange dots: a simulated tree generated by the program of Boettcher and Paczuski.

Figure 8 shows log-log plots of distribution of normalized ultrametric distances for two computer-generated trees. The ultrametric distances were binned by log 2 and there are 100 bins. A histogram was obtained and converted to the current dot-line style.

The dots represent the probability density in each bin. The blue and orange dots represent the Markov and simulation tree, respectively.

A straight line in a log-log graph corresponds to a power law and its slope is the exponent of the power law. In the distribution of ultrametric distances, a less steep slope (i.e. a smaller absolute value, less negative value) indicates long-range correlations and strong memory effects. For a steep slope, the exponent of the power law is more negative, and the tail falls off rapidly, resulting in correlations in shorter range and weaker memories. We found the equations of the two fitting lines are: $y_o = 4x_o^{-0.6}$ for the Markov tree, and $y_r = 6.7 \times 10^3 x_r^{-1.85}$ for the simulated tree.

It meets our expectation that the simulation gives a straight line while the Markov tree as a memoryless tree should not give a power law. On the log-log graph, a Poisson distribution should fall off exponentially and fall earlier than a power law. However, the Markov does not show this pattern as we expected for a memoryless process. The branch lengths were created from a Poisson distribution, but due to the interaction between lattice sites, a factor of "surface roughness" was added into their final distribution and disrupted the Poisson pattern (Krug & Spohn, 1988).



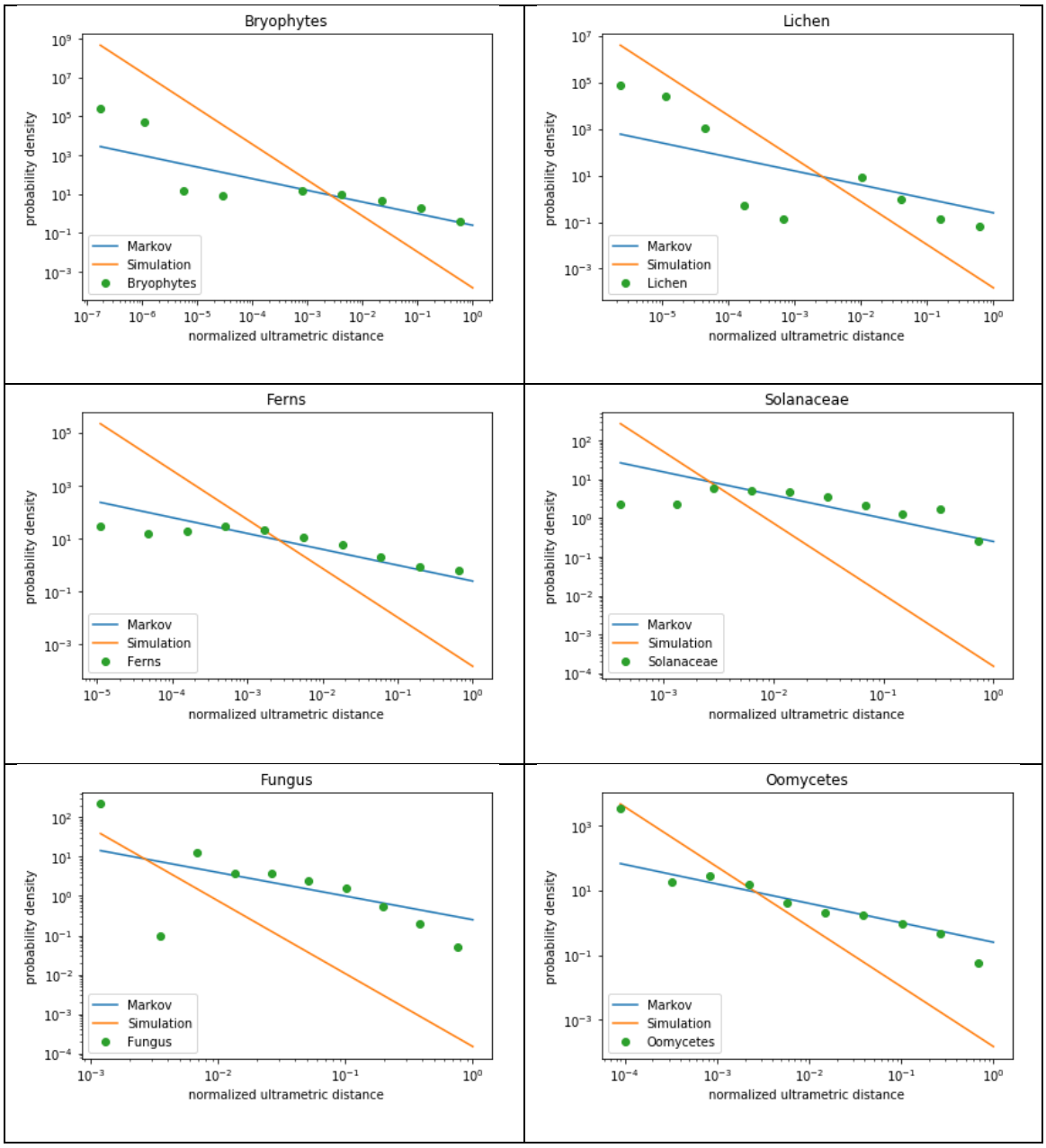


Figure 9. Log-log plots of the distribution of normalized ultrametric distances for eight natural trees that show a power law pattern. From top left to bottom right: (1) Armillaria. (2) Apiaceae. (3) Brophytes. (4) Lichen. (5) Ferns. (6) Solanaceae. (7) Fungus. (8) Oomycetes.

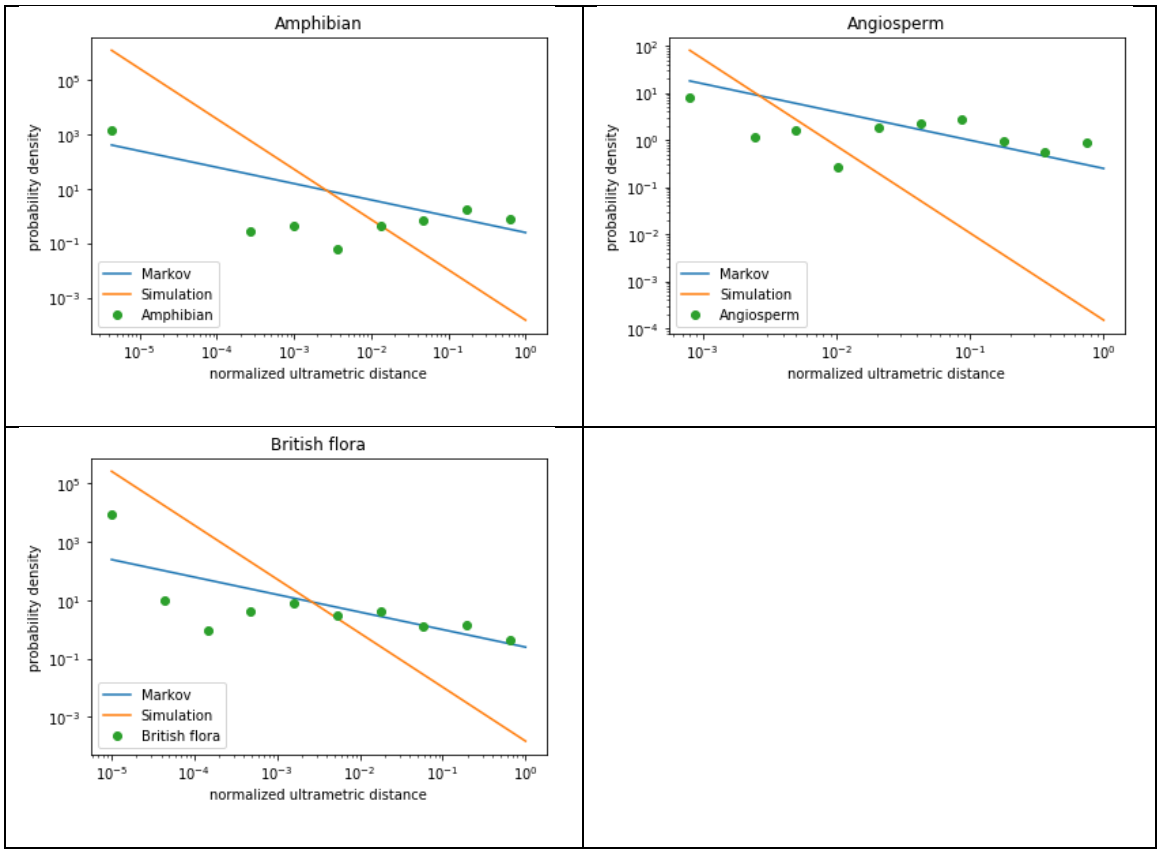


Figure 10. Distribution of normalized ultrametric distances for three natural trees with strongest memories (relatively flat slopes). (1) Amphibian. (2) Angiosperm. (3) British flora.

We found that scale-invariance is present in all trees. The distribution of data points more or less follows a straight line rather than being scattered around, which shows a power law pattern. A property of power laws is the scale-invariance. It implies that relatively simple and consistent rules of evolution govern the diversification, although the driving forces may be different for each tree because of the different slops.

We also found that a long-term memory effect exists in varying degrees. In evolutionary terms, the ultrametric distances represent how much time has elapsed between a lowest common ancestor and two consecutive speciation events. It speaks to

the timing information that the waiting time of later speciation events depends on the time of ancestral events (memory). The three trees in figure 10 have relatively flatter slope, indicating very strong memories. The Lichen tree in figure 9.4 has one of the weakest memories. Armillria, Apiaceae, Lichen, and Fungus (figure 9.1, 2, 4, 7) possess partial resemblance to the fitting line of simulated tree. For example, although the data points of Lichen do not form a perfect line and can be divided into three sectors, each sector is parallel to the line of simulated tree. Their evolutionary process may be described by the simulation model. In the introduction, we briefly discussed that the model incorporates the idea of fitness and as a result of this model, the evolutionary process of the simulated tree was described by punctuated equilibrium. Therefore, punctuated equilibrium rather than phyletic gradualism may play a more important role in evolution of the four trees here.

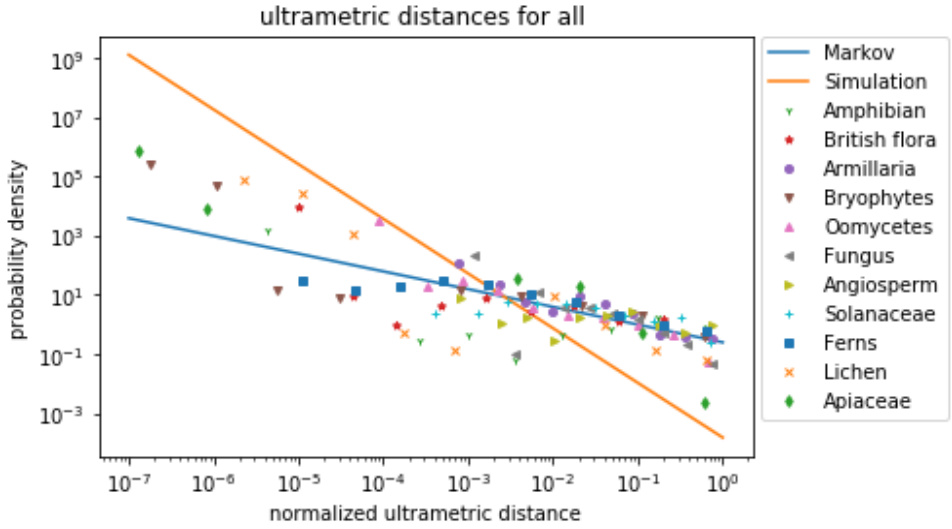


Figure 11. Ultrametric distances for all trees in one plot. The blue line is the fitting line of the Markov tree; the orange line is the fitting line of the simulated tree; the dots correspond to data points of each reconstructed phylogenetic tree, respectively.

Trees in figure 9.3, 5, 6, 8 show proximity to the Markov tree. For example, almost all data points of fern tree in figure 9.5 follow the Markov line, making it the closest to the Markov process among trees in our study. It is contrary to our expectation and other studies using different measurements that natural trees are distinguishable from the Markov tree (Herrada et al., 2008; Xue et al., 2020). Moreover, as shown in figure 11, the overall trend of phylogenetic trees at large distances falls on the Markov line. It implies that when species difference is disregarded and the phylogenetic tree is talked as a non-specific and collective term, the overall evolutionary process seems to be memoryless. However, considering evolution in this way is unjustified by evolutionary theories as specific environmental factors deeply influence the evolution of a particular species.

We have detected the presence of universal scaling and long-term memory effects in the 11 trees we have studied. However, the exponents of natural trees are closer to the Markov tree. It opens room for future study. First, the finding that the topology and branch lengths distribution of some trees are closer to the Markov tree may arise from two possible reasons: first, their evolutionary processes do resemble the Markov process; second, the ratio of clade sizes and ultrametric distances may not be good indicators because other studies using different measurements found results distinguishable from their null models. For the first reason, future study could look deeper into the biological details of each species to identify certain factors that causes the particular memoryless process. We could also borrow the measuring tools from other studies to see whether our natural trees are still close to the null model. For the second reason, we could use our methods to analyze trees that are proved by other studies to be distinguishable from the

Markov process. Also, more theoretical works are needed to analyze why the clade size ratio and ultrametric distances are not good indicators.

Second, there are risks that the comparison between the natural trees and the models is not justified. In the two models, there are two types of leaves: those that extinct in the past and those that live in the present. The majority of leaves are extinct. However, some natural trees have the same mixture of leaves, but some natural trees such as fern only have current-living leaves. Extinction events are not recorded in the latter. If the data of all extinct leaves in the model were removed to ensure a fair comparison between a model and a natural tree with no extinction, the results would not be statistically significant.

Third, we didn't explore the relation between the two indicators (the distribution of ultrametric distances and the ratio of clade size), which may be valuable to study in the future. Lastly, a potential systematic bias is present, as 10 out of 11 reconstructed trees are plants and fungi and only the amphibian tree belongs to animals.

4. References

Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science*, *16*(1), 23–34. https://doi.org/10.1214/ss/998929474

Aldous, D. J., Krikun, M. A., & Popovic, L. (2011). Five Statistical Questions about the Tree of Life. *Systematic Biology*, *60*(3), 318–328. https://doi.org/10.1093/sysbio/syr008

Alexander Pyron, R., & Wiens, J. J. (2011). A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, *61*(2), 543–583. https://doi.org/10.1016/j.ympev.2011.06.012

Agapow, P.-M., & Purvis, A. (2002). Power of Eight Tree Shape Statistics to Detect Nonrandom Diversification: A Comparison by Simulation of Two Models of Cladogenesis. Systematic Biology, 51(6), 866–872. https://doi.org/10.1080/10635150290102564

Bainard, J. D., Newmaster, S. G., & Budke, J. M. (2020). Genome size and endopolyploidy evolution across the moss phylogeny. *Annals of Botany*, *125*(4), 543–555. https://doi.org/10.1093/aob/mcz194

Bak, P., & Sneppen, K. (1993). Punctuated equilibrium and criticality in a simple model of evolution. *Physical Review Letters*, *71*(24), 4083–4086. https://doi.org/10.1103/PhysRevLett.71.4083

Banasiak, Ł., Piwczyński, M., Uliński, T., Downie, S. R., Watson, M. F., Shakya, B., & Spalik, K. (2013). Dispersal patterns in space and time: A case study of Apiaceae subfamily Apioideae. *Journal of Biogeography*, *40*(7), 1324–1335. https://doi.org/10.1111/jbi.12071

Bernardi, M., Angielczyk, K. D., Mitchell, J. S., & Ruta, M. (2016). Phylogenetic Stability, Tree Shape, and Character Compatibility: A Case Study Using Early Tetrapods. Systematic Biology, 65(5), 737–758. https://doi.org/10.1093/sysbio/syw049

Billera, L. J., Holmes, S. P., & Vogtmann, K. (2001). Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics*, *27*(4), 733–767. https://doi.org/10.1006/aama.2001.0759

Blum, M. G. B., & François, O. (2006). Which Random Processes Describe the Tree of Life? A Large-Scale Study of Phylogenetic Tree Imbalance. *Systematic Biology*, *55*(4), 685–691. https://doi.org/10.1080/10635150600889625

Boettcher, S., & Paczuski, M. (1996). Ultrametricity and Memory in a Solvable Model of Self-Organized Criticality. *Physical Review E*, *54*(2), 1082–1095. https://doi.org/10.1103/PhysRevE.54.1082

Burke, D. J., Carrino-Kyker, S. R., & Burns, J. H. (2019). Is it climate or chemistry? Soil fungal communities respond to soil nutrients in a multi-year high-resolution analysis. *Ecosphere*, *10*(10), e02896. https://doi.org/10.1002/ecs2.2896

Coetzee, M. P. A., Wingfield, B. D., Bloomer, P., Ridley, G. S., & Wingfield, M. J. (2003). Molecular identification and phylogeny of Armillaria isolates from South

America and Indo-Malaysia. *Mycologia*, *95*(2), 285–293. https://doi.org/10.1080/15572536.2004.11833113

Colijn, C., & Plazzotta, G. (2018). A Metric on Phylogenetic Tree Shapes. *Systematic Biology*, *67*(1), 113–126. https://doi.org/10.1093/sysbio/syx046

Davies, T. J., Allen, A. P., Borda-de-Água, L., Regetz, J., & Melián, C. J. (2011). Neutral Biodiversity Theory Can Explain the Imbalance of Phylogenetic Trees but Not the Tempo of Their Diversification. *Evolution*, *65*(7), 1841–1850. https://doi.org/10.1111/j.1558-5646.2011.01265.x

Gascuel, F., Ferrière, R., Aguilée, R., & Lambert, A. (2015). How Ecology and Landscape Dynamics Shape Phylogenetic Trees. *Systematic Biology*, *64*(4), 590–607. https://doi.org/10.1093/sysbio/syv014

Heard, S. B. (1996). Patterns in Phylogenetic Tree Balance with Variable and Evolving Speciation Rates. *Evolution*, *50*(6), 2141–2148. https://doi.org/10.1111/j.1558-5646.1996.tb03604.x

Hermant, M., Hennion, F., Bartish, I. V., Yguel, B., & Prinzing, A. (2012). Disparate relatives: Life histories vary more in genera occupying intermediate environments. *Perspectives in Plant Ecology, Evolution and Systematics*, *14*(4), 283–301. https://doi.org/10.1016/j.ppees.2012.02.001

Hernandez-Garcia, E., Tugrul, M., Herrada, E. A., Eguiluz, V. M., & Klemm, K. (2010). Simple models for scaling in phylogenetic trees. *International Journal of Bifurcation and Chaos*, *20*(03), 805–811. https://doi.org/10.1142/S0218127410026095

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*(7), e2757. https://doi.org/10.1371/journal.pone.0002757

Jones, G. R. (2011). Tree Models for Macroevolution and Phylogenetic Analysis. Systematic Biology, 60(6), 735–746. https://doi.org/10.1093/sysbio/syr086

Kozyrev, S. V. (2011). Methods and applications of ultrametric and p-adic analysis: From wavelet theory to biophysics. Proceedings of *the Steklov Institute of Mathematics*, *274*(1), 1. https://doi.org/10.1134/S0081543811070017

Krug, J., & Spohn, H. (1988). Universality classes for deterministic surface growth. *Physics Review A*, *38*(8), 4271-4283. https://link.aps.org/doi/10.1103/PhysRevA.38.4271

Leavitt, S. D., Kraichak, E., Nelsen, M. P., Altermann, S., Divakar, P. K., Alors, D., Esslinger, T. L., Crespo, A., & Lumbsch, T. (2015). Fungal specificity and selectivity for algae play a major role in determining lichen partnerships across diverse ecogeographic regions in the lichen-forming family Parmeliaceae (Ascomycota). *Molecular Ecology*, *24*(14), 3779–3797. https://doi.org/10.1111/mec.13271

Lim, J., Crawley, M. J., Vere, N. D., Rich, T., & Savolainen, V. (2014). A phylogenetic analysis of the British flora sheds light on the evolutionary and ecological factors driving plant invasions. *Ecology and Evolution*, *4*(22), 4258–4269. https://doi.org/10.1002/ece3.1274

Liu, P., Gould, M., & Colijn, C. (2020). Polynomial Phylogenetic Analysis of Tree Shapes. *BioRxiv*, 2020.02.10.942367. https://doi.org/10.1101/2020.02.10.942367

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). NEXUS: An Extensible File Format for Systematic Information. *Systematic Biology*, *46*, 32.

Matsen, F. A. (2006). A Geometric Approach to Tree Shape Statistics. *Systematic Biology*, *55*(4), 652–661. https://doi.org/10.1080/10635150600889617

Mooers, A. O., & Heard, S. B. (1997). Inferring Evolutionary Process from Phylogenetic Tree Shape. *The Quarterly Review of Biology*, *72*(1), 31–54. https://doi.org/10.1086/419657

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, *17*(4), 508–525. https://doi.org/10.1111/ele.12251

Mossel, E., & Steel, M. (2004). A phase transition for a random cluster model on phylogenetic trees. *Mathematical Biosciences*, *187*(2), 189–203. https://doi.org/10.1016/j.mbs.2003.10.004

Rammal, R., Toulouse, G., & Virasoro, M. A. (1986). Ultrametricity for physicists. *Reviews of Modern Physics*, *58*(3), 765–788. https://doi.org/10.1103/RevModPhys.58.765

Robideau, G. P., de Cock, A. W. A. M., Coffey, M. D., Voglmayr, H., Brouwer, H., Bala, K., Chitty, D. W., Désaulniers, N., Eggertson, Q. A., Gachon, C. M. M., Hu, C.-H., Küpper, F. C., Rintoul, T. L., Sarhan, E., Verstappen, E. C. P., Zhang, Y., Bonants, P. J. M., Ristaino, J. B., & Lévesque, C. A. (2011). DNA barcoding of oomycetes

with cytochrome c oxidase subunit I and internal transcribed spacer. *Molecular Ecology Resources*, *11*(6), 1002–1011. https://doi.org/10.1111/j.1755-0998.2011.03041.x

Särkinen, T., Bohs, L., Olmstead, R. G., & Knapp, S. (2013). A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): A dated 1000-tip tree. *BMC Evolutionary Biology*, *13*(1), 214. https://doi.org/10.1186/1471-2148-13-214

Testo, W., & Sundue, M. (2016). A 4000-species dataset provides new insight into the evolution of ferns. *Molecular Phylogenetics and Evolution*, *105*, 200–211. https://doi.org/10.1016/j.ympev.2016.09.003

Vandewalle, N., & Ausloos, M. (1997). Construction and properties of fractal trees with tunable dimension: The interplay of geometry and physics. *Physical Review E*, *55*(1), 94–98. https://doi.org/10.1103/PhysRevE.55.94

Xue, C., Liu, Z., & Goldenfeld, N. (2020). Scale-invariant topology and bursty branching of evolutionary trees emerge from niche construction. *Proceedings of the National Academy of Sciences*, *117*(14), 7879–7887. https://doi.org/10.1073/pnas.1915088117

Ultrametric space. (2021). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Ultrametric_space&oldid=1001637780

5. Appendix

Table 1. A list of phylogenetic trees downloaded from TreeBASE.

No.	Category	Tree Type	# taxa	Tree ID	Reference
1	Amphibian	Species Tree	2872	Tr48025	Pyron & Wiens (2011)
2	Armillaria (Honey Fungus)	Species Tree	1124	Tr96228	Coetzee et al. (2003)
3	Angiosperm (flowering plant)	Species Tree	1284	Tr60915	Hermant et al. (2012)
4	Apiaceae (Umbellifers)	Species Tree	1194	Tr91597	Banasiak et al. (2013)
5	British flora	Species Tree	1653	Tr91679	Lim et al. (2014)
6	Bryophytes	Species Tree	2214	Tr121673	Bainard et al. (2019)
7	Ferns	Species Tree	4007	Tr100612	Testo & Sundue (2016)
8	Fungus	Species Tree	1127	Tr110559	Burke et al. (2019)
9	Lichen	Species Tree	2356	Tr88158	Leavitt et al. (2015)
10	Oomycetes	Species Tree	1205	Tr46272	Robideau et al. (2011)
11	Solanaceae (nightshade)	Species Tree	1076	Tr100305	Särkinen et al. (2014)

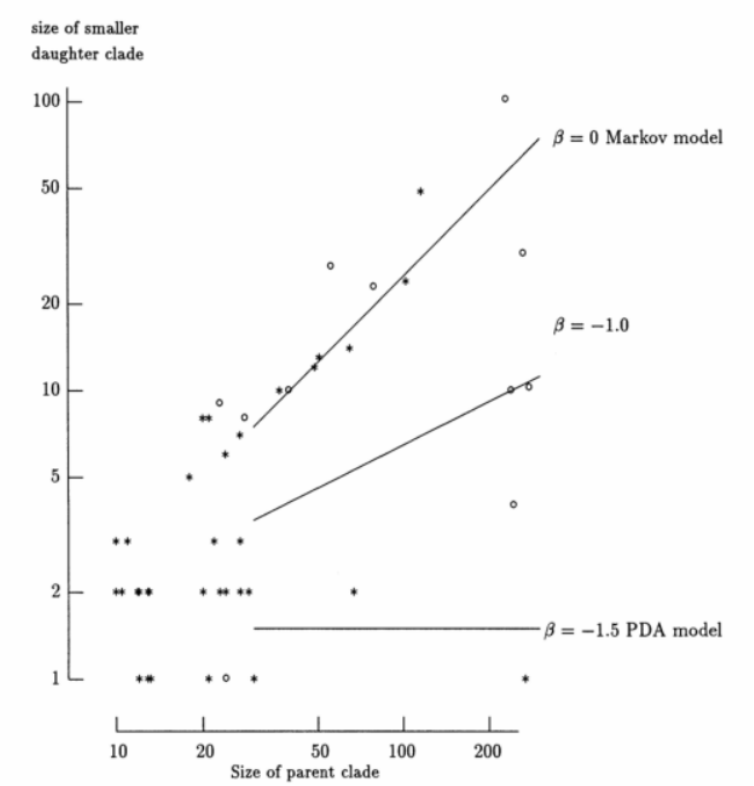
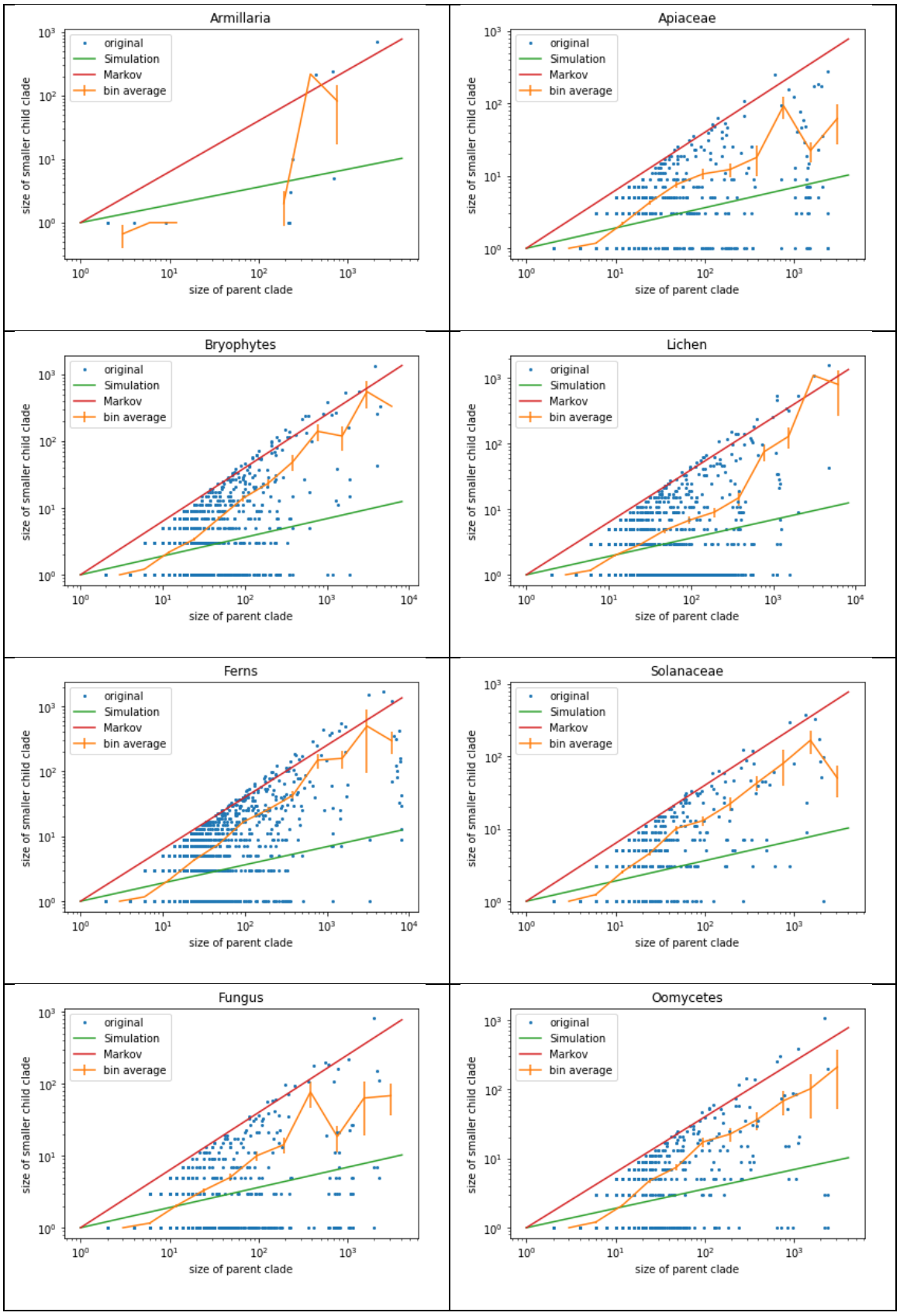


FIG. 4. *Splits in the tree of Harrington* (1980).

Figure 12. The size of smaller daughter clade versus size of parent clade in Aldous (2001). The graph is copied here for a quick reference. The three lines are the simulations corresponding to three values of beta in the beta-splitting family.


Armillaria
Apiaceae
Bryophytes
Lichen
Ferns
Solanaceae
Fungus
Oomycetes
original
Simulation
Markov
bin average
size of smaller child clade
size of parent clade

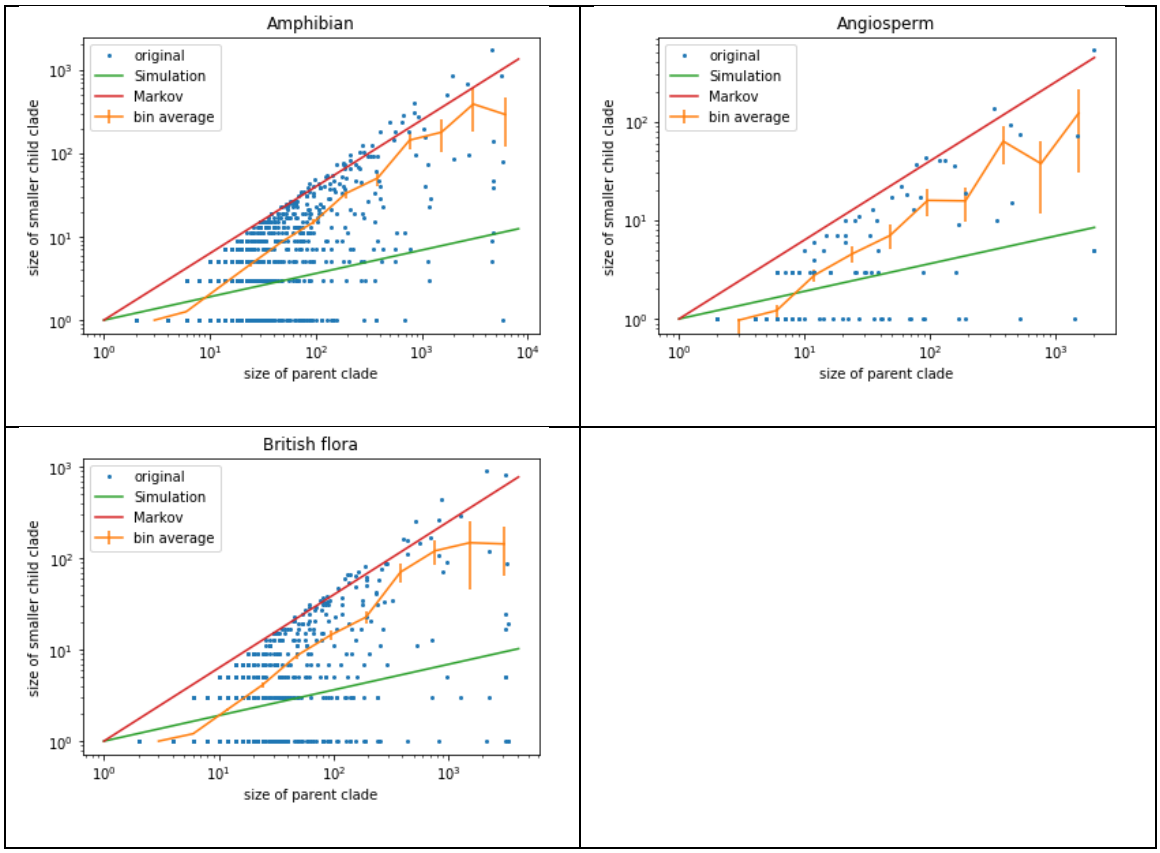


Figure 13. The log-log plot of the size of smaller child clade over the size of parent clade for all phylogenetic trees. From top left to bottom right: (1) Armillaria. (2) Apiaceae. (3) Brophytes. (4) Lichen. (5) Ferns. (6) Solanaceae. (7) Fungus. (8) Oomycetes. (9) Amphibian. (10) Angiosperm. (11) British flora.