On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited

Michael G.B. Blum†‡ Olivier François†

March 30, 2005

†TIMC-TIMB, Faculté de Médecine, F38706 La Tronche cedex, France
‡Laboratoire Ecologie, Systématique et Evolution UMR 8079, Bâtiment 360, Université Paris-Sud, 91405 Orsay Cedex, France

Abstract

We investigate the distribution of statistical measures of tree imbalance in large phylogenies. More specifically, we study normalized versions of the Sackin's index and the number of subtrees of given sizes. Using the connection with structures from theoretical computer science, we provide precise description for the limiting distribution under the null hypothesis of Yule trees. Corrected p-values are then computed, and the statistical power of these statistics for testing the Yule model against a model of biased speciation is evaluated from simulations. As an illustration, the tests are applied to the HIV1 reconstructed phylogeny.

1 Introduction

Phylogenetic trees are widely used in biology to represent evolutionary relationships between species (Nei and Kumar, 2000). A second kind of application is to the study of cladogenesis. In this case, the shape of a phylogenetic tree conveys useful information about the process by which it has grown (Harvey et al, 1996). It may reflect for example the fingerprint of the rates of species formation and extinction. Measuring the degree of imbalance or asymmetry of a tree topology may therefore provide support for the hypothesis that species have different potential for speciation.

Several statistics have been introduced for assessing the level of asymmetry of a tree. These statistics are often used to test whether the tree topology differs significantly from a null model in which the rates of speciation are constant among species (Kirkpatrick and Slatkin, 1993; Mooers and Heard, 1997). The null model is commonly known as the equal-rates Markov model or *Yule model* (Yule, 1924). In the Yule model on rooted trees, each external branch has an equal probability of splitting (Athreya and Ney, 1972).

Among imbalance statistics, the most classical are Sackin's index (Sackin, 1972; Shao and Sokal, 1990) and Colless' index (Colless, 1982). Sackin's index is the average path length from a tip to the root of the tree. Colless' index inspects the internal nodes, partitioning the tips that descend from them into groups of sizes r and s, and computes the sum of absolute values $|r - s|$ for all nodes. Colless's index is often renormalized for giving the value one to the totally pectinate tree. More recently, McKenzie and Steel (2000) proposed to count the number of *cherries*, ie the number of pairs of leaves that are

adjacent to a common ancestor node.

The power of five imbalance statistics have been evaluated by Kirkpatrick and Slatkin (1993) who concluded that Sackin and Colless statistics were among the most powerful with respect to an alternative model of biased speciation. These works were extended by Agapow and Purvis (2002) regarding more biologically motivated alternative models. These authors reached similar conclusions.

In this article, we consider fully dichotomous trees (binary trees) with n leaves (and $(n-1)$ internal nodes). We focus on Sackin's index and its connection to a series of statistics that are similar to the number of cherries. Our emphasis is on the fact that all these measures are relevant to the same empirical distribution, namely the distribution of the size of subtrees. Sackin's index is connected to the expectation of the empirical distribution while cherries merely correspond to subtrees of size two.

The main results presented in this article can be summarized as follows. First, we give a description of the limiting distribution of the Sackin's index for large n. The limiting distribution is non Gaussian, and can be defined as the solution of a functional fixed-point equation. Second, we extend McKenzie and Steel's result for the number of cherries in a Yule tree to the number (or frequencies) of subtrees of size larger than two.

These extensions are based on a link with existing results in theoretical computer science regarding binary search trees. These trees appear as formal representation for *divide-and-conquer* algorithms (Rösler, 1992; Hwang and Neininger, 2002). We exploit the one-to-one correspondence between binary

search trees and Yule trees in order to describe the asymptotic distributions of the Sackin's Index and the size of subtrees.

In addition, we propose a new statistic based of the computation of a ℓ_1 distance between the empirical distribution of the number of subtrees and the theoretical distribution. The power of all statistics to reject the null hypothesis of a Yule tree against a model of biased speciation are then evaluated.

The article is organized as follows. Section 2 presents the asymptotic theory for the Sackin's index and the number of subtrees of a given size. Section 3 describes statistical tests with proper corrections based on the theory and their five percent confidence intervals. Section 4 evaluates the statistical power of these tests based on simulation studies. An example of application to the HIV tree is discussed in Section 5.

2 Theory

2.1 Sackin's statistic

Sackin's statistic is one of the oldest measure that summarizes the shape of a tree (Sackin, 1972; Shao and Sokal, 1990). It adds the number of internal nodes between each leaf of the tree and the root to form the following index

$$I_s^n = \sum_{i=1}^{n} N_i$$

where the sum runs over the n leaves of the tree and N_i is the number of internal nodes crossed in the path from i to the root (including the root). An equivalent formulation of I_s^n is by counting the number of leaves under

each internal nodes

$$I_s^n = \sum_{j=1}^{n-1} \tilde{N}_j$$

where $\tilde{N}_j$ is the number of leaves that descend from the ancestor j. This is a well-known result in systematic biology that the expectation of I_s^n under the Yule model is of order $2n \log n$ (eg, Kirkpatrick and Slatkin, 1993)

$$E[I_s^n] = 2n \sum_{j=2}^{n} 1/j \, .$$

The variance is more complex, but it can be estimated by noticing the analogy with a classical problem in theoretical computer science. This analogy is a crucial step in defining the proper correction for indices of large phylogenies. Let us explain this briefly. Binary trees are data structures often encountered in computer science, more specifically in connection with *divide and conquer* algorithms. To each Yule tree corresponds a binary search tree in a unique manner (see Aldous, 2001). Using this one-to-one correspondence, Sackin's statistic is equal to the number of comparisons used by the quicksort algorithm to sort a random input (eg, Rösler, 1992). This can also be seen directly because the Sackin's index is involved in a stochastic recurrence equation

$$I_s^n = I_s^J + I_s^{n-J} + n \tag{1}$$

where J is a uniform random variable over the subset $\{1, \ldots, n-1\}$. The recurrence equation is obtained by splitting the tree at the root into two sister clades (the left and right subtrees). Conditional on J, the values I_s^J and I_s^{n-J} are independent random variables which correspond to the indices of the left and right subtrees.

Standard computations lead to the result that the variance of I_s^n is of order $\sigma^2 n^2$ where σ^2 is independent of n (eg, Hwang and Neininger, 2002). In addition, the normalized Sackin's index (to which we refer in the sequel) can be defined as

$$I_s = \frac{I_s^n - E[I_s^n]}{n}. \tag{2}$$

The normalized index I_s converges in distribution as the number of leaves n grows to infinity. According to Rösler (1992), the limit X satisfies a (functional) fixed-point equation of the following type

$$X = UX + (1-U)X^* + 2U\log U + 2(1-U)\log(1-U) + 1 \tag{3}$$

where X, X^*, U are independent random variables, X and X^* are identically distributed, U is uniformly distributed over the inverval $(0,1)$, and the equality holds for distributions. Using equation (3), the variance of the limiting distribution can be computed in an exact way

$$\sigma^2 = 7 - \frac{2\pi^2}{3}.$$

2.2 Number of cherries

McKenzie and Steel (2000) considered a simple and easily computed statistic for evaluating tree shape: the number of cherries of the tree. A *cherry* is a pair of leaves that are adjacent to a common ancestor node. The authors analysed the distribution of this statistic under the Yule model. They obtained exact formulae for the mean and variance of the number of cherries, and showed that this distribution is asymptotically normal as the number of leaves grows

to infinity. More specifically, if we denote by C_n the number of cherries in a tree of size n, their results can be summarized as follows

$$E[C_n] = \frac{n}{3}$$

and

$$\mathrm{Var}[C_n] = \frac{2n}{45}\,.$$

Using an argument based on extended Polya urns, McKenzie and Steel obtained that

$$\frac{C_n - n/3}{\sqrt{2n/45}} \to \mathcal{N}(0,1)\,.$$

2.3 The number of subtrees of fixed size

Sackin's statistic and the number of cherries are two distinct aspects of the same distribution: the number of leaves under a randomly chosen node. Let Z_n denote this number. On the one hand, Z_n is connected to Sackin index by the fact that

$$Z_n = \tilde{N}_J,$$

where J is a uniform random variable over that subset $\{1, \ldots, n-1\}$ (recall that $\tilde{N}_j$ is the number of leaves that descend from the ancestor j). Given the tree structure T, one actually has

$$E[Z_n \mid T] = \frac{1}{n-1}\sum_{j=1}^{n-1} \tilde{N}_j = \frac{I_s^n}{n-1}\,.$$

On the other hand, the empirical frequency of cherries in a Yule tree $f_n(2) = C_n/(n-1)$ is an unbiased estimator of the probability of the event $(Z_n = 2)$.

The distribution of Z_n has been described by Blum and François (2004) using results from coalescent theory.

Theorem 1 *(Blum and François, 2004) Let $n \geq 2$ and Z_n be the number of individuals in a uniformly chosen random clade of a Yule tree with n leaves. We have*

$$p_n(z) = \mathrm{P}(Z_n = z) = \frac{n}{(n-1)} \frac{2}{z(z+1)}, \quad z = 2, \ldots, n-1,$$

and

$$p_n(n) = \mathrm{P}(Z_n = n) = \frac{1}{n-1}.$$

By the above remarks, we can check that $(n-1)Z_n$ has the same average value that Sackin's statistics (both are equal to the average complexity of the quicksort algorithm). In addition, we find that the frequencies of subtrees of size 2 is connected the number of cherries as follows

$$f_n(2) = \frac{C_n}{n-1} = \frac{1}{n-1} \sum_{j=1}^{n-1} \mathbb{1}_{(N_j=2)}$$

Taking expectation, this leads to

$$E[f_n(2)] = E[\frac{1}{n-1} \sum_{j=1}^{n-1} \mathbb{1}_{(N_j=2)}] = \mathrm{P}(N_J = 2) = \frac{n}{3(n-1)}, \tag{4}$$

and we can recover the fact that $E[C_n] = n/3$.

Normality of frequencies of subset sizes For a Yule tree T with n leaves, let $f_n(z)$ denote the frequencies of subtrees of size z in the tree ($z \geq 2$). Equation (4) tells that the frequency $f_n(2)$ is an unbiased estimator of the

probability $p_n(2)$. The same argument applies to proving that $f_n(z)$ is an unbiased estimator of the probability $p_n(z)$ for all $z \geq 2$. In addition, we obtain that the limiting distribution of $f_n(z)$ is Gaussian as n goes to infinity.

Theorem 2 *Let $z \geq 2$. The empirical probabilities $f_n(z)$ have variances of order $1/n$*

$$\mathrm{Var}[f_n(z)] = \frac{\sigma^2 n}{(n-1)^2} \sim \frac{\sigma^2}{n}.$$

In addition, the following convergence in distribution holds

$$\sqrt{n}\,(f_n(z) - p_n(z)) \to \mathcal{N}(0, \sigma^2(z)), \quad \text{as } n \to \infty$$

where, for all $z \geq 2$,

$$\sigma^2(z) = \frac{2(z-1)(4z^2-3z-4)}{z(2z+1)(2z-1)(z+1)^2}.$$

Proof. Let $X_n^z = (n-1)f_n(z)$ denote the number of families of size z. The case $z = 2$ is a direct consequence of McKenzie and Steel's results (2000) because $X_n^2 = C_n$ is equal to the number of cherries in a coalescent or Yule tree. In addition, $\sigma^2(2)$ is equal to $2/45$. Let us give a sketch of proof for the general result. For $z \geq 3$, the random variable X_n^z can be involved into a quicksort-like recurrence equation (see Hwang and Neininger, 2002)

$$X_n = X_J + X_{n-J}^* + t_n \tag{5}$$

where J is uniformly taken from the set $\{1, \ldots, n-1\}$ and the *toll* function t_n is equal to

$$t_n = \delta_{n,z}$$

where $\delta_{n,z}$ denotes the Kronecker symbol. In this equation, X and X^* are independent copies which correspond to the values obtained for the left and right subtrees after a split at the root of the tree. The formal expression of the variance of X_n^z was computed using equation (5) and elementary programming in MAPLE. For $n \geq 2z+1$, we obtained that $\mathrm{Var}[X_n^z] = \sigma^2(z)n$ with $\sigma^2(z)$ given by the Theorem. The final result follows from Hwang and Neininger (2002). ∎

Comments. A rather direct proof of Theorem 2 can be found in (Devroye, 1991). Devroye states the result for binary search trees. To obtain the above result, it must be modified according to the one-to-one correspondence between binary search trees and Yule trees. A binary search tree with $(n-1)$ nodes corresponds identically to a Yule tree with n leaves (see Aldous (2001)). Another proof for the mean and variance formulae was given by Rosenberg (2004) by purely combinatorial techniques.

For all $z \geq 2$ and n sufficiently large, five percent level confidence intervals for frequencies are given by

$$-1.96\sigma(z)\frac{\sqrt{n}}{n-1} < f_n(z) < +1.96\sigma(z)\frac{\sqrt{n}}{n-1}$$

The accuracy of the approximation depends on z. For $z \leq 10$, simulation evidences show that Gaussian distributions provide good fit to the empirical distributions of $f_n(z)$ as soon as $n \geq 30$.

2.4 New indices

To conclude this section, we propose a new statistic based on the comparison of the empirical and theoretical distributions of the number of subtrees under the Yule model. This index is defined as a weighted ℓ_1 distance

$$D = \sum_{z=2}^{n} z \, |f_n(z) - p_n(z)|$$

where the sum runs over the all possible subset sizes under an arbitrary node.

In spirit, the metric D is comparable to the Sackin's index, giving importance to the apparition of abnormally large number of leaves under the nodes close to the root. However it has the advantage of providing a one-sided test, whereas Sackin index provides a two-sided test.

3 Distribution of indices

In this section, we estimate the quantiles of the distributions of the normalized Sackin's index and the distance index D for samples of various sizes.

Experimental design In order to estimate the quantiles of the distributions $10,000$ Yule trees were simulated. All simulations were performed under the object-oriented R language, which provide facilities for manipulating tree data (R Development Core Team, 2003). Empirical cumulative distribution functions were computed and displayed in Figures 1 and 2. Approximate values of the empirical quantiles can easily be determined from these graphical representations.

Normalized Sackin Normalized values of the Sackin index have been considered earlier by Kirkpatrick and Slatkin (1993). However, these authors proposed the value 1.96 for statistical significance at the five percent level (Gaussian approximation). We propose corrections that account for the fact that the distribution of the normalized Sackin's index is non Gaussian. For $n = 30$ taxa, the five percent rejection area can be described as

$$I_s < -0.72 \quad \text{or} \quad I_s > 1.24$$

For $n = 100$, the five percent rejection area is slightly different

$$I_s < -0.87 \quad \text{or} \quad I_s > 1.43.$$

The limiting distribution was computed numerically according to equation 1 (using a method developed by Tan and Hadjicostas (1995)), and good agreement with the empirical distribution was noticed for $n = 100$.

Distance D Regarding the distance D, no normalization is available theoretically. Empirical quantiles for samples of size $n = 30, 100, 200$ can be deduced from the Figure 2. The test is significant at the five percent level if $D > 5.10$ for $n = 30$, $D > 8.04$ for $n = 100$ and $D > 9.65$ for $n = 200$. Simulation results show that D should be corrected to $\tilde{D} = D - 2.27 \log(n) + 3.62$ (almost perfect log-linear fit). The five percent rejection area is $\tilde{D} > 1.25$ for large n.

4 Statistical power

4.1 Biased speciation model

A basic issue regarding the power of these different statistics is which one is the most sensitive to a departure from the Yule model. This question is also relevant to coalescent models and population genetics where departure from neutrality is an important step in detecting the evolutionary pressures acting on a sample of genes.

The power of statistics depends on the alternative hypothesis and there are many possible choices that may produce imbalanced trees. For instance, Pinelis introduced a family of models that emcompass the traditional hypotheses about tree-biased speciation (Pinelis, 2003). Kirkpatrick and Slatkin (1993) evaluated the power of five tree shape statistics encompassing the Sackin and Colless indices at three tree sizes (10, 20 and 40 species). They generated imbalanced phylogenies by making the instantaneous rates of speciation of every pair of sister lineages differ by a constant factor. They concluded that the Sackin and Colless indices performed well (except for the smallest trees).

Agapow and Purvis considered other processes of non random speciation in which the rate of speciation in a lineage evolves inpendently of the rates in other lineages and is directed toward greater biological relevance (see Agapow and Purvis, 2002). Among 8 studied statistics, they found that the Sackin and Colless indices performed well for models of trait evolution. These indices were a little less powerful when applied to a model of age-dependent rates.

Biased speciation model In this study, we used an alternative model of biased speciation which is similar to the one used by Kirkpatrick and Slatkin. Assume that the speciation rate of a specific lineage is equal to r ($0 \leq r \leq 1$). When a species with speciation rate r splits, one of its descendent species is given the rate pr and the other is given the speciation rate $(1-p)r$ where p is fixed for the entire tree. These rates are effective until the daughter species themselves speciate. Values of p close to 0 or 1 yield very imbalanced trees while values around 0.5 lead to over-balanced phylogenies.

We simulated this model for different numbers of species $n = 30, 100, 200$ and different values of p. The most interesting values are around $p = 0.12 - 0.15$ where it may sometimes be difficult to detect the imbalance visually. Type two errors β were calculed from $10,000$ independent Monte Carlo repetitions and the type one error α was fixed at the level of 5 percent.

Results The results are reported in Tables 1-3. The performance of the statistics $f_n(z)$ to detect imbalance were weak for small ($n = 30$) phylogenies. They were sightly better for larger trees $n = 100 - 200$. Among subtrees, counting the number of cherries appeared to be the most efficient way of detecting departure from the Yule model. Overall, $f_n(z)$ and C_n showed very low power for all z, and we would not recommend their use for testing imbalance.

Sackin statistics I_s and the D distance were very powerful as concerned high desequilibrium, ie $p = 0.05$ for $n = 30$, and $p \leq 0.1$ for $n = 100, 200$. In this situation, I_s was sightly more powerful than D. When imbalance is

less evident $p = 0.125 - 15$, the performances of both indices decreased. In this situation, we typically obtained 86 % of errors for D while this ratio was equal to 92 % for I_s. The decrease of power was thus slower for D. The last rows of Table 1-3 show that D was unable to detect over-balanced trees while I_s performed rather well in this regard (due to the two-sided test).

5 Example

This section analyzes a dataset taken from the literature (Yusim et al., 2001). Several authors attempted to infer historical features of the acquired immune deficiency syndrome (AIDS) using human immunodeficiency virus type 1 (HIV-1) sequences. There are three distinctive form of HIV-1 (M,O,N). Group M contains the viruses which cause the global HIV pandemic and appear to have arisen in Central Africa during the last 100 years (Korber et al., 2000). Vidal et al. (2000) investigated the genetic diversity of HIV-1 group M in this region by obtaining viral gene sequences in 1997 from 197 infected individuals living in the Democratic Republic of Congo. Yusim et al. (2001) used a maximum likelihood approach to estimate a phylogeny for this large data set, and it is this phylogeny that we use here. The phylogeny is availabe from the R package *ape* (Paradis et al., 2004).

Korber et al. (2000) estimated the time since the most recent common ancestor of the HIV-1 thanks to this tree, and this was compared with a coalescent approach by Yusim et al. (2001). These works were based on the assumption of a molecular clock, or a constant rate of evolution among each lineage. Rambaut et al. (2001) noticed that it is unlikely that this HIV-

1 data set has been evolving according to this hypothesis. Therefore, the goodness-of-fit of Yule or coalescent (essentially the same topological model) to this dataset has to be tested.

To assess the fit to a Yule model, we computed the normalized Sackin index, the empirical distribution of subset sizes, and the weighted ℓ_1 distance D. The Sackin's index was equal to $I_s = 0.82$ ($\mathrm{P}(I_s > 0.82) = 0.10$). The empirical distribution $f_n(z)$ of subset sizes is given in Table 4. The test provided by theorem 2

$$|f_n(z) - p_n(z)| > 1.96\sigma^2(z)\sqrt{n/(n-1)}$$

was non-significant for all $z \leq 10$. The distance D was equal to $D = 9.37$ ($\tilde{D} = 1.05$) and the p-value was equal $\mathrm{P}(\tilde{D} > 1.05) \approx 0.08$. We could not conclude to the rejection of the Yule or coalescent models.

To go further, we remark that imbalance might be detected at any level of the tree, and we considered cutting the branches of the tree which were far from the root. Doing so, we kept only the "old" internal branches that corresponded to the 30 oldest ancestors. Under the null hypothesis for the $n = 192$ sequences, the pruned tree should also be compatible with a Yule model. In this case, the Sackin's index was equal to $I_s = 1.21$ ($\mathrm{P}(I_s > 1.21) = 0.03$). The empirical distribution $f_n(z)$ of subset sizes is given in Table 5. The test provided by Theorem 2 was significant for $z = 2$ (cherries). The distance D was equal to $D = 5.17$ and the p-value was $\mathrm{P}(D > 5.17) \approx 0.04$. These results probably indicate a change in the evolutionary rate during the evolution which had more impact on cladogenesis during the early expansion of the virus.

Acknowledgments. We are grateful to an anonymous referee for the time she/he has spent reading our article and for her/his useful comments.

References

[1] Agapow P-M., A. Purvis. 2002. Power of Eight Tree Shape Statistics to Detect Nonrandom Diversification: A Comparison by Simulation of Two Models of Cladogenesis. Syst. Biol. 51(6):866-872.

[2] Aldous, D.J. 1991. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.* 1, 228-266.

[3] Aldous, D.J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science*, **16** 23-34.

[4] Athreya K.B., P.E. Ney. 1972. Branching Processes, Springer, Berlin.

[5] Blum M.G.B., O. Francois. 2004. External branch length and minimal clade size under the neutral coalescent model. Submitted.

[6] Devroye L. 1991. Limit laws for local counters in random binary search trees. *Random Structure and Algorithms* 2, 303-315.

[7] Harvey, P.H., A.J. Leigh Brown, J. Meynard Smith, S. Nee. 1996. *New uses for new phylogenies*. Oxford: Oxford University Press.

[8] Hwang H-K., R. Neininger. 2002. Phase change of limit laws in the quicksort recurrence under varying toll functions. SIAM J. Comput. Vol. 31, 6, pp. 1687-1722.

[9] Kirkpatrick M., M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree, Evolution 47 (4), 1171-1181.

[10] Korber B, M Muldoon, J Theiler, F Gao, R Gupta, R Lapedes, B Hahn, S Wolinsky, T Battacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains, Science 288, 1789-1796.

[11] McKenzie A., M. Steel. 2000. Distributions of cherries for two models of trees. Mathematical Biosciences, 164 pp. 81-92.

[12] Mooers A., S.B. Heard. 1997. Inferring evolutionary process from phylogenetic tree shape, Quart. Rev. Biol. 72 (1) 31-54.

[13] Nei M., S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

[14] Paradis E., Claude J., K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 20: 289-90.

[15] Pinelis I. 2003. Evolutionary models of phylogenetic trees. Proc. R. Soc. Lond. B 270: 1425-1431.

[16] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2003.

[17] Rambaut A, Robertson DL, Pybus OG, Peeters M, and Holmes EC. Human immunodeficiency virusPhylogeny and the origin of HIV-1. Nature 2001 410:1047-1048.

[18] Rosenberg N. 2004. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Preprint submitted.

[19] Rösler U. 1992. A limit theorem for quicksort. RAIRO Inform. Theor. Appl., 25:85-100.

[20] Rösler U. 2001. The analysis of divide and conquer algorithms, Algorithmica 29: 238-261.

[21] Sackin, M. J. 1972. "Good" and "bad" phenograms. Systematic Zoology 21:225-226.

[22] Shao K., and R. R. Sokal. 1990. Tree balance. Systematic Zoology, 39, 266-276.

[23] Tan H.K., and P. Hadjicostas. 1995. Some properties of a limiting distributions of quicksort. Statist. Probab. Lett., 25, 87-94.

[24] Vidal, N., M. Peeters, C. Mulanga-Kayeba, N. Nzilambi, D. Robertson, W. Ilunga, H. Sema, K. Tshimanga, B. Bongo, E. Delaporte. 2000. Unprecedented degree of HIV-1 group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J Virol. 74, 10, 498-507.

[25] Yule, G.U. 1924. A mathematical theory of evolution, based on the conclusions of Dr J.C. Willis. *Philos. Trans. Roy. Soc. London Ser. B*, **213** 1924, 21-87.

[26] Yusim K., M. Peeters, O.G. Phybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler, B. Korber. 2001. Phil. Trans. R. Soc. Lond. B 356, 855-866.

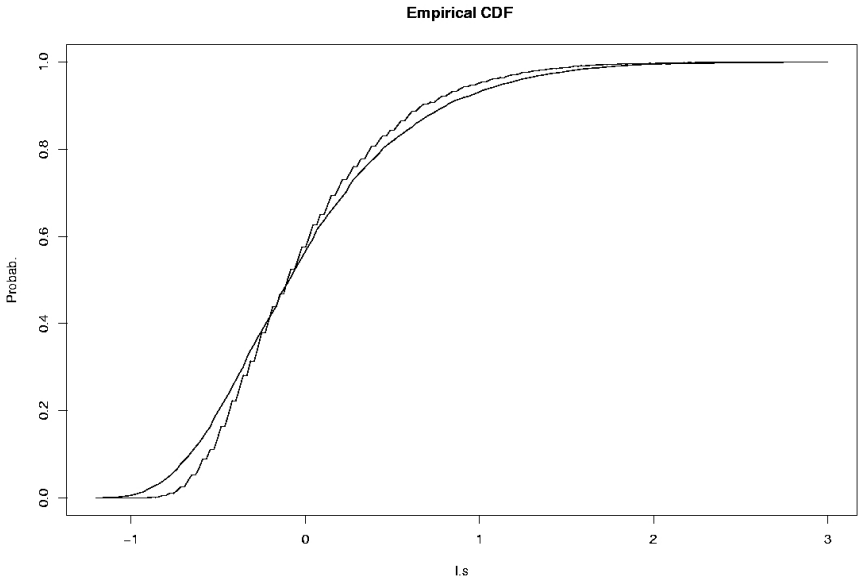

Figure 1: *Empirical cumulative distribution functions of the normalized Sackin's index under the Yule model for random trees of sizes* $n = 30$ *and* $n = 100$ *(smooth line). The cdfs were computed from 10,000 trees*

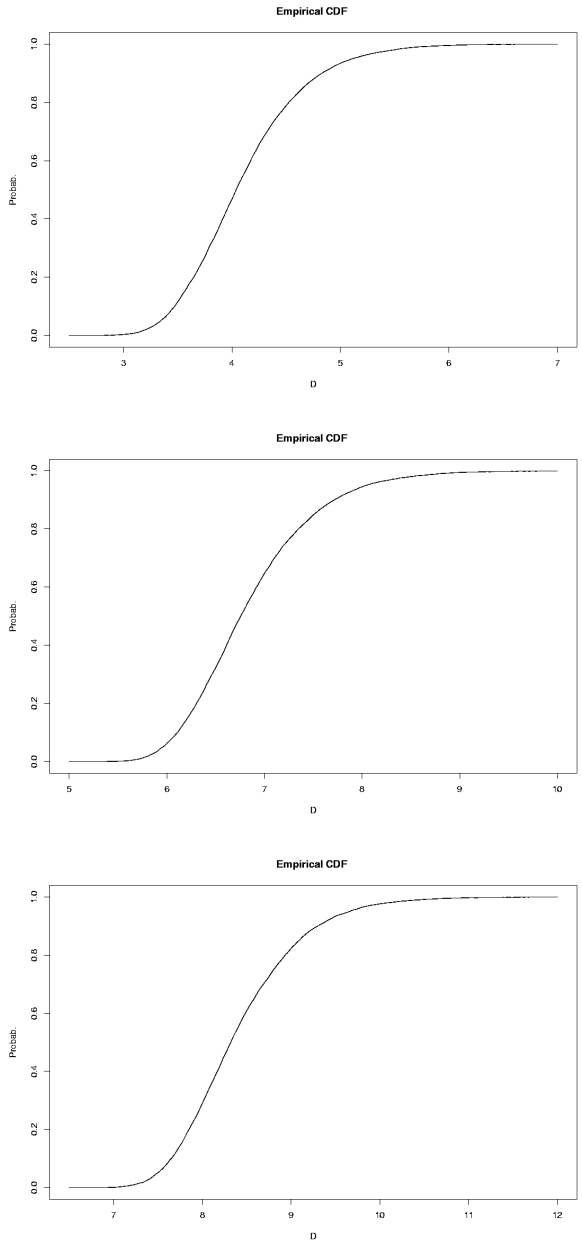

Figure 2: *Empirical cumulative distribution functions of the D statistic under the Yule model for random trees of sizes $n = 30$, $n = 100$, and $n = 200$.*

$n = 30$ p	D	I_s	$f_n(2)$	$f_n(3)$	$f_n(4)$	$f_n(5)$	$f_n(6)$	$f_n(7)$	$f_n(8)$
$p = 0.05$	0.02	0.005	0.79	0.80	0.95	1	–	–	–
$p = 0.1$	0.42	0.36	0.95	0.96	0.98	1	–	–	–
$p = 0.125$	0.69	0.72	0.96	0.96	0.99	1	–	–	–
$p = 0.15$	0.86	0.92	0.97	0.96	0.98	1	–	–	–
$p = 0.25$	0.99	0.98	0.97	0.95	0.98	0.99	0.99	0.96	0.99
$p = 0.4$	0.99	0.35	0.93	0.94	0.97	1	0.99	0.96	0.99
$p = 0.5$	1	0.11	0.92	0.93	0.97	1	1	0.95	0.98

Table 1: *Type two error β for the alternative hypothesis H_1 of biased speciation with parameters $n = 30$, and $p = 0.05 - 0.5$ computed from 10,000 repetitions. The type one error is $\alpha = 0.05$.*

$n = 100$ p	D	I_s	$f_n(2)$	$f_n(3)$	$f_n(4)$	$f_n(5)$	$f_n(6)$	$f_n(7)$	$f_n(8)$	$f_n(9)$	$f_n(10)$
$p = 0.05$	0	0	0.21	0.96	0.99	0.99	0.98	0.97	0.90	–	–
$p = 0.1$	0.01	0.00	0.55	0.88	0.99	0.98	0.97	0.94	0.94	–	–
$p = 0.125$	0.24	0.27	0.78	0.90	0.98	0.99	0.98	0.94	0.93	0.92	–
$p = 0.15$	0.77	0.90	0.91	0.92	0.98	0.99	0.98	0.96	0.94	0.96	–
$p = 0.25$	0.99	0.98	0.89	0.95	0.97	0.97	0.97	0.94	–	–	–
$p = 0.4$	0.99	0	0.63	0.92	0.96	0.96	0.96	0.96	0.92	–	–
$p = 0.5$	1	0	0.57	0.91	0.95	0.96	0.95	0.95	0.92	–	–

Table 2: *Type two error β for the alternative hypothesis H_1 of biased speciation with parameters $n = 100$, and $p = 0.05 - 0.5$ computed from 10,000 repetitions. The type one error is $\alpha = 0.05$.*

$n = 200$ p	D	I_s	$f_n(2)$	$f_n(3)$	$f_n(4)$	$f_n(5)$	$f_n(6)$	$f_n(7)$	$f_n(8)$	$f_n(9)$	$f_n(10)$
$p = 0.05$	0.02	0.01	0.001	0.75	0.99	0.90	0.87	0.88	0.92	0.92	0.96
$p = 0.1$	0.42	0.36	0.49	0.77	0.88	0.97	0.99	0.96	0.93	0.84	0.82
$p = 0.125$	0.69	0.71	0.83	0.92	0.91	0.94	0.98	0.98	0.97	0.92	0.90
$p = 0.15$	0.86	0.92	0.93	0.95	0.94	0.94	0.98	0.98	0.98	0.97	0.95
$p = 0.25$	1	0.92	0.81	0.93	0.93	0.93	0.97	0.97	0.97	0.95	0.95
$p = 0.4$	1	0.0	0.32	0.86	0.92	0.92	0.95	0.96	0.96	0.96	0.95
$p = 0.5$	1	0.0	0.24	0.85	0.91	0.92	0.95	0.96	0.96	0.96	0.95

Table 3: *Type two error β for the alternative hypothesis H_1 of biased speciation with parameters $n = 200$, and $p = 0.05 - 0.5$ computed from 10,000 repetitions. The type one error is $\alpha = 0.05$.*

	$z=2$	$z=3$	$z=4$	$z=5$	$z=6$	$z=7$
theoretical	.335	.167	.100	.067	.047	.035
empirical	.312	.140	.114	.057	.057	.036

Table 4: *Theoretical and empirical distribution $f_n(z)$ of the size of subsets in the HIV-1 phylogeny group M. The number of sequences is $n = 192$, and z denotes the size of subsets.*

	$z=2$	$z=3$	$z=4$	$z=5$	$z=6$	$z=7$
theoretical	.344	.172	.103	.068	.049	.036
empirical	.241	.172	.103	.068	.068	.068

Table 5: *Theoretical and empirical distribution $f_{30}(z)$ of the size of subsets in the HIV-1 phylogeny group M. The tree was pruned to keep the 30 oldest internal nodes (the top of the tree).*