DIMACS Series in Discrete Mathematics
and Theoretical Computer Science
Volume 61, 2003

The Size of a Maximum Agreement Subtree for Random Binary Trees

David Bryant, Andy McKenzie, and Mike Steel

ABSTRACT. In computational biology, a common way to compare two rooted trees that classify the same set L of labelled leaves is to determine the largest subset of L on which the two trees agree. In this paper we consider the size of this quantity if one or both trees are generated randomly, according to two simple null models. We obtain analytical bounds, as well as providing some simulation results that suggest a power law similar to the related problem of determining the length of the longest increasing sequence in a random permutation.

1. Introduction

Hierarchical relationships in evolutionary biology and linguistics are often represented by rooted binary trees with leaves labelled by the species under study. Frequently two trees that classify the same set of species will differ if they have been obtained from different data sets, or by using different methods. In these situations it is useful to know what is the largest subset of species on which the two trees agree. Informally, a "maximum agreement subtree" for a pair T_1, T_2 of trees, each having the same set L of labelled leaves, is a tree t with leaves lying in a largest possible subset of L, so that the t can be embedded in both T_1 and T_2. A maximum agreement subtree (MAST) offers a way of summarising what information two (or more) trees have in common. The concept was introduced by [**6**] and [**9**].

The first polynomial-time algorithms for computing a MAST for two trees were developed independently by [**8**] and [**13**]. Later it was shown that computing a MAST for three or more trees is an NP-hard problem, though solvable in polynomial time if a degree bound is placed on at least one of the input trees [**2**]. The algorithm

2000 *Mathematics Subject Classification.* Primary 92B10, 60J20; Secondary 92D10, 68R10.
Key words and phrases. trees, phylogeny, maximum agreement subtree, distributions on trees.
We thank the New Zealand Marsden Fund (UOC-MIS-005) for supporting this research.

of [3] for computing MAST trees has been implemented as part of the widely used phylogenetic analysis software PAUP* [15].

In this paper we are not concerned with algorithms for finding a MAST. Rather, we investigate the probability that the size of a MAST for two binary trees exceeds any given value when one or both of the trees are randomly generated. We also consider the expected size of a MAST under the two models we consider.

These questions are relevant to biology when comparing evolutionary trees for the same set of species that have been constructed from two quite different types of data. It may be suspected that one or both of the data sets contain no phylogenetic information; for example with DNA sequence data this can occur if the sites are sufficiently saturated due to high mutation rates. In this case one or both of the reconstructed trees are essentially random and so it is useful to know how much agreement one should expect between the two trees purely by chance.

This question is also closely related to a classical problem in combinatorics. Suppose we generate a permutation $x_1, \ldots, x_n$ on the numbers $1, \ldots, n$ uniformly at random and ask for the longest monotone increasing subsequence $x_{i(1)}, \ldots, x_{i(s)}$. Then, the ratio $(s/\sqrt{n})$ converges in probability to 2 as n tends to infinity ([4], p.369).

This result has some bearing on the MAST problem. Let us call a rooted binary tree a *comb* phylogeny if every non-leaf vertex is unlabelled and is adjacent to at least one labelled leaf. From the result on permutations it follows that, for two comb phylogenies generated uniformly at random on the same set of n labelled leaves, the size of the MAST divided by $\sqrt{n}$ converges in probability to 2 as n tends to infinity.

However comb phylogenies are rather special, and it is interesting to inquire as to whether a power law for the size of maximum agreement subtrees holds in general for more interesting probability distributions on leaf-labelled binary trees. Our analytic results show that a $\sqrt{n}$ behaviour is an upper bound for the expected size of the MAST for an underlying uniform distribution. Simulations also support a power law similar to $\sqrt{n}$ as a lower bound for the expected size.

We end this section by noting that the deterministic minimum size of a MAST of two binary trees has been investigated by [7]. These authors found that the relative depths of the two trees was an important factor in setting lower bounds on the size of a MAST.

2. Terminology

In this paper a *phylogeny on* L is a rooted binary tree T consisting of a *root vertex* of degree two, unlabelled *interior vertices* of degree three, and *leaf* vertices of degree one that are labelled bijectively from the set L.

We will let $L(T) = L$ denote the set of labelled leaves of T, the *leaf set of* T. Usually we will take $L = [n] = \{1, \ldots, n\}$.

Phylogenies are widely used to represent evolutionary relationships in biology, where they are sometimes also referred to as "rooted binary phylogenetic trees". The leaf set L corresponds to the extant species under study, while the remaining interior vertices correspond to speciation events.

Let T be a phylogeny with leaf set $L(T) = [n]$, and let S be a subset of $[n]$ of size at least two. The *induced phylogeny* $T|_S$ is the phylogeny on S obtained by taking the minimal subtree of T that connects all the leaves in S, and then supressing any resulting non-root vertices of degree 2. For example, in Figure 1, we have $t = T|_{\{1,2,3,5\}}$. We will generally use the lower case letter t to denote induced phylogenies.

We denote the set of phylogenies on leaf set $[n]$ and the size of this set by $|RB(n)|$. It is a classical and well known result that $|RB(n)| = (2n-3)!!$ where

$$(2n-3)!! = (2n-3) \times (2n-5) \times \cdots \times 5 \times 3 = \frac{(2n-2)!}{(n-1)!2^{n-1}}.$$

Let T and T' be two phylogenies on the leaf set $[n]$ and let S be a subset of $[n]$ of size $s > 1$. Then a phylogeny t on S is an *agreement subtree* for the pair T, T' if $T|_S = T'|_S = t$.

If, in addition, t has the maximum number of leaves amongst all agreement subtrees for the pair T, T' we say that t is a *maximum agreement subtree* (or MAST) for T, T'. The number of leaves in any MAST is called the *size* of the MAST, and is denoted $M(T, T')$. That is,

$$M(T, T') = \max\{|S| : S \subseteq L, T|_S = T'|_S\}$$

See Figure 1 for an example of a pair of phylogenies that have a MAST of size four. Note that a MAST may not be unique; moreover the number of MASTs for two phylogenies on a leaf set of size n can grow exponentially with n [**11**].

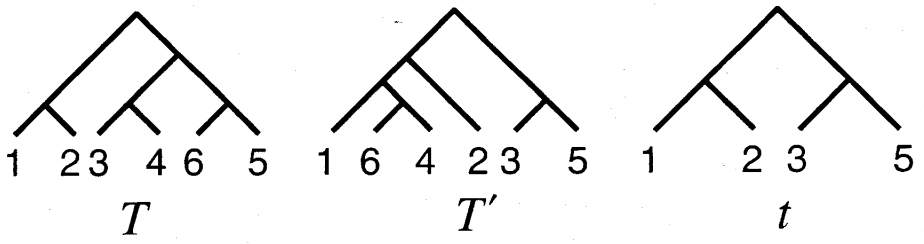

FIGURE 1. The phylogeny t is a MAST for the phylogenies T and T'. We have $M(T, T') = 4$.

3. The Yule-Harding and Uniform Models

There has been considerable interest in investigating models for the generation of phylogenies. Such models can be useful for testing speciation hypotheses, or as

null models for Bayesian approaches to phylogeny. We now discuss two natural models that have been used for generating phylogenies.

The *Yule-Harding* model (also called the "Markov model") can be defined in several apparently different ways. One definition, which emphasises the link with the speciation process, is in terms of edge addition. One starts with a tree on two randomly selected species from L. To the tree so-far constructed, a leaf v is selected uniformly at random, its incident edge is subdivided, and the resulting degree-two vertex is made adjacent to a new leaf, labelled uniformly at random by one of the remaining elements of L that are not already a leaf label of the tree. This process is repeated to give a tree with $L(T) = L$. The underlying rationale for this process are the assumptions that speciation is instantaneous, always occurs as bifurcations, is independent across lineages, and that the probability of speciation is the same for all lineages at any given time [**12**].

A second model is the *Uniform model* in which equal probability is assigned to each possible phylogeny on the leaf set L. Thus, under this model the probability of a particular phylogeny is $1/|RB(n)|$, where $n = |L|$. This model is sometimes referred to as the "proportional-to-distinguishable-arrangements (PDA) model". As with the Yule-Harding model, the Uniform model can also be realized by a certain speciation scenario (for details see [**14**]).

For any stochastic model that generates phylogenies, if we are given a phylogeny T with a leaf set L of size n, we will let $\mathbb{P}_n[T]$ denote the probability that a randomly generated phylogeny on L is T.

As explicitly noted by [**1**], the Yule-Harding and Uniform models on phylogenies satisfy the following two properties:

- **Exchangeability** If T and T' are phylogenies on $[n]$ and T' can be obtained from T by permuting the elements of $[n]$ then $\mathbb{P}_n[T] = \mathbb{P}_n[T']$.
- **Sampling consistency.** For a phylogeny T on $[n]$, and $s < n$ we have

$$\mathbb{P}_n[T|_{\{1,2,\ldots,s\}} = t\,] = \mathbb{P}_s[t]\,, \tag{3.1}$$

  where t is a phylogeny on $[s]$.

In Figure 2 we show the results of some simulations that suggest a power law relationship between the expected value of $M(T, T')$ and n under the Uniform and Yule-Harding models. All simulated expected values were based on 1000 randomly generated trees. The simulations also suggested that $M(T, T')$ has an approximately normal distribution about its mean.

4. Upper Bounds

We now derive an upper bound for the probability that two randomly generated trees have a MAST of size greater than or equal to any given value s. In the two subsequent sections we determine this upper bound explicitly for the Uniform model and recursively for the Yule-Harding model.

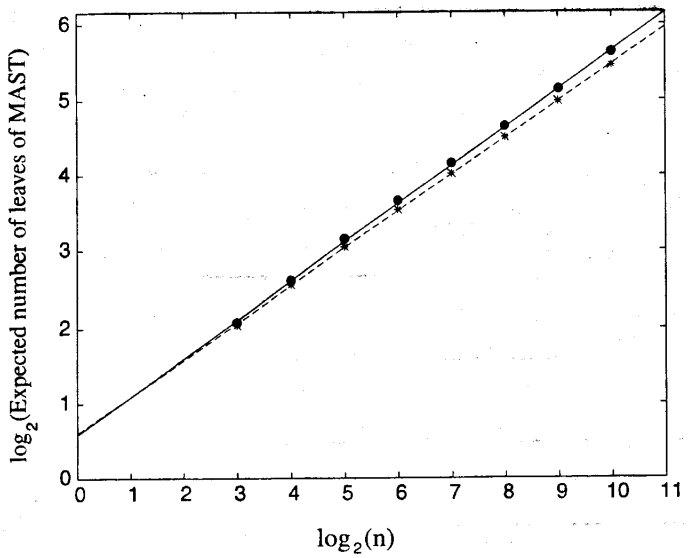

FIGURE 2. A log-log scaled plot of the expected size of a MAST vs number of leaves on each tree (n). Simulated values for the Yule-Harding ($\bullet$) and Uniform ($*$) models. The least-squares line of best fit for the Yule-Harding model data (–) has a slope 0.506 and an intersect of 0.595. For the Uniform model the least-squares line (--) has slope 0.487 and intersect 0.603.

Given phylogenies T and T' on $[n]$, and a subset S of $[n]$ let

$$X_S = \begin{cases} 1, & \text{if } T|_S = T'|_S; \\ 0, & \text{otherwise .} \end{cases} \tag{4.1}$$

The number of agreement subtrees with s leaves for T and T' is then counted by

$$X^{(s)} = \sum_{S \subseteq [n]:|S|=s} X_S \, . \tag{4.2}$$

Now suppose that phylogenies T and T' on $[n]$ are randomly generated according to some model. Let $\psi_{n,s}$ be the expected number of agreement subtrees of T and T' of size s.

LEMMA 4.1. *Suppose that phylogenies T and T' on a leaf set L of size n are randomly generated under a model that satisfies exchangeability and sampling consistency. Then,*

$$\mathbb{P}[M(T,T') \geq s] \leq \psi_{n,s} = \binom{n}{s} \sum_{t \in RB(s)} \mathbb{P}_s[t]^2 \, .$$

TABLE 1. The simulated expected number of leaves for the MAST of two randomly generated trees on n leaves under the Yule-Harding and Uniform models. The simulated upper five percent value represents the smallest value of $s : \mathbb{P}[M(T,T') \geq s] \leq 0.05$. All simulated values are based on 1000 randomly generated trees. The bound upper five percent value is the value of s such that the upper bound on $\mathbb{P}[M(T,T') \geq s]$ in Lemma 4.1 is less than or equal to 0.05.

	Yule-Harding model			Uniform model		
Number of leaves (n)	Simulated $\mathbb{E}[M(T,T')]$	Simulated upper 5%	Bound upper 5%	Simulated $\mathbb{E}[M(T,T')]$	Simulated upper 5%	Bound upper 5%
8	4.2	6	6	4.1	6	6
16	6.1	8	9	5.9	8	9
32	8.9	11	13	8.2	11	12
64	12.6	16	18	11.6	14	17
128	17.8	21	25	16.2	19	23
256	25.1	29	35	22.7	26	32
512	35.3	40	50	31.6	36	45
1024	49.4	54	70	43.8	49	63

Proof. The event $\{M(T,T') \geq s\}$ is equivalent to the event $\{X^{(s)} \geq 1\}$ so

$$\mathbb{P}[M(T,T') \geq s] = \mathbb{P}[X^{(s)} \geq 1] \leq \mathbb{E}[X^{(s)}] = \psi_{n,s} = \sum_{S \subseteq [n]:|S|=s} \mathbb{E}[X_S]$$

$$= \sum_{S \subseteq [n]:|S|=s} \mathbb{P}[X_S = 1] = \binom{n}{s} \mathbb{P}[X_{\{1,2,\ldots,s\}} = 1],$$

where the last equality is by exchangeability. Now,

$$\mathbb{P}[X_{\{1,2,\ldots,s\}} = 1] = \mathbb{P}_n[T\,|_{\{1,2,\ldots,s\}} = T'|_{\{1,2,\ldots,s\}}]$$

$$= \sum_{t \in RB(s)} \mathbb{P}_n[T\,|_{\{1,2,\ldots,s\}} = t \text{ and } T'|_{\{1,2,\ldots,s\}} = t]$$

$$= \sum_{t \in RB(s)} \mathbb{P}_n[T|_{\{1,2,\ldots,s\}} = t]^2 = \sum_{t \in RB(s)} \mathbb{P}_s[t]^2,$$

where the last equality follows from the sampling consistency property. Upon substituting back for this term, we obtain the upper bound as stated in the lemma. □

Table 1 shows that the bound described by Lemma 4.1 provides an approximate estimate of the tail of the distribution of $M(T,T')$. By comparing the estimate of the smallest value of s for which $\mathbb{P}[M(T,T') \geq s] \leq 0.05$ obtained from the analytical bound, with values obtained from simulations, we see that the former tend to overestimate the latter by between 0 and 30 percent in the range shown.

4.1. Uniform Model. We derive some analytic bounds for the Uniform model when one or both phylogenies are randomly generated. In this section we will suppose that T is either a fixed phylogeny on $[n]$, or is generated according to any distribution on phylogenies on $[n]$ (regardless of whether or not this distribution satisfies sampling consistency or exchangeability). We will further suppose that T'

is independently generated by the Uniform model for phylogenies on $[n]$. We let M_n denote the random variable $M(T, T')$.

PROPOSITION 4.2. Suppose T and T' are phylogenies on $[n]$ with T fixed or random, and T' independently generated according to the Uniform model.

$$\mathbb{P}[M_n \geq s] \leq \psi_{n,s} = \binom{n}{s} \frac{1}{(2s-3)!!}.$$

Proof. Following the proof of Lemma 4.1 we have

$$\mathbb{P}[M_n \geq s] \leq \sum_{S \subseteq [n]:|S|=s} \mathbb{P}[X_S = 1].$$

Now, regardless of whether T is fixed, or selected according to any distribution, we have

$$\mathbb{P}[X_S = 1] = \sum_t \mathbb{P}[T'|_S = t]\mathbb{P}[T|_S = t] = \sum_t \frac{1}{|RB(s)|}\mathbb{P}[T|_S = t] = \frac{1}{|RB(s)|}$$

and the result now follows. □

As we show in the following theorem, the asymptotic behaviour of the quantity $\psi_{n,s}$ depends on the limiting value of the ratio $\frac{n}{s^2}$. As usual we write $f(m) \sim g(m)$ to denote that $\lim_{m\to\infty} \frac{f(m)}{g(m)} = 1$. It is also useful to have the concept of exponential convergence to zero, and we write $f(m) \to_{\exp} 0$ if, for some constant $\epsilon \in (0,1)$, and some integer m_0, $f(m) < \epsilon^m$ for $m > m_0$.

THEOREM 4.3. *Suppose T and T' are phylogenies on $[n]$ with T fixed or random, and T' independently generated according to the Uniform model.*

(i) *The expected number of agreement subtrees of size s for the pair T, T' is given by*

$$\psi_{n,s} = \sqrt{\frac{s}{\pi}} \left(\frac{ne^2}{2s^2}\right)^s Y(n,s)\theta(s)$$

where $Y(n,s) = \prod_{i=1}^{s-1}(1 - \frac{i}{n})$ and where the function θ satisfies the condition $\theta(s) \sim 1$.

(ii) *Let*

$$f_\lambda^+(s) = \max\{\psi_{n,s} : n \leq s^2/\lambda^2\}; f_\lambda^-(s) = \inf\{\psi_{n,s} : n \geq s^2/\lambda^2\}.$$

If $\lambda > \frac{e}{\sqrt{2}}$, $f_\lambda^+(s) \to_{\exp} 0$, while if $\lambda < \frac{e}{\sqrt{2}}$, $f_\lambda^-(s) \to \infty$ as $s \to \infty$.

(iii) *For $\lambda > \frac{e}{\sqrt{2}}$, let $g(n) = \mathbb{P}[M_n \geq \lambda\sqrt{n}]$. Then $g(n) \to_{\exp} 0$.*

(iv) *For any $\lambda > \frac{e}{\sqrt{2}}$ there exists a value m so that, for all $n \geq m$,*

$$\mathbb{E}[M_n] \leq \lambda\sqrt{n}.$$

Proof. Let

$$\theta(s) = \frac{\sqrt{\frac{\pi}{s}}\left(\frac{2s^2}{e^2}\right)^s}{s!(2s-3)!!}$$

Since $(2s-3)!! = \frac{(2s-2)!}{(s-1)!2^{s-1}}$, we can apply using Stirling's approximation (see [5]), which states that $s! \sim \sqrt{2\pi}s^{s+1/2}e^{-s}$, to deduce that $\theta(s) \sim 1$. Now,

$$\psi_{n,s} = \frac{n!}{s!(n-s)!(2s-3)!!} = \frac{n^s Y(n,s)}{s!(2s-3)!!}$$

and it can be checked that the right hand side of the this equation now equals $\sqrt{\frac{s}{\pi}}\left(\frac{ne^2}{2s^2}\right)^s Y(n,s)\theta(s)$ which establishes part (i).

For part (ii), the result for f_λ^+ follows from part (i) by observing that $Y(n,s) \leq 1$, and that $\sqrt{s}\alpha^s \to_{\exp} 0$ for $\alpha = \frac{e^2}{2\lambda^2} < 1$. The result for f_λ^- also follows from part (i) by noting that, provided $n > c_1 s^2$ for some positive constant c_1, then $Y(n,s) > c_2 > 0$ for an associated constant c_2, and $\lim_{s\to\infty}\sqrt{s}\beta^s = \infty$ for $\beta > 1$.

For part (iii) and part (iv) let us select $\epsilon > 0$ sufficiently small so that $\lambda - \epsilon > \frac{e}{\sqrt{2}}$. Then we may select, for each $n > \frac{1}{\epsilon^2}$, an integer s_n such that $(\lambda-\epsilon)\sqrt{n} \leq s_n \leq \lambda\sqrt{n}$. Now,

$$\mathbb{P}[M_n \geq \lambda\sqrt{n}] \leq \mathbb{P}[M_n \geq s_n] \leq \psi_{n,s_n}, \tag{4.3}$$

and since $\psi_{n,s_n} \leq \max\{\psi_{k,s_n} : k \leq s_n^2/(\lambda-\epsilon)^2\}$ it follows from (4.3) that

$$\mathbb{P}[M_n \geq \lambda\sqrt{n}] \leq f^+_{(\lambda-\epsilon)}(s_n). \tag{4.4}$$

By part (ii) the right hand side of (4.4) converges to zero exponentially with s_n and thereby with n. This establishes part (iii).

For part (iv) note that, since M_n is bounded above by n, the following inequality holds for any positive real value of $s \leq n$,

$$\mathbb{E}[M_n] \leq s\mathbb{P}[M_n < s] + n\mathbb{P}[M_n \geq s] = (n-s)\mathbb{P}[M_n \geq s] + s, \tag{4.5}$$

where the second equality holds because $\mathbb{P}[M_n < s] = 1 - \mathbb{P}[M_n \geq s]$. With ϵ chosen as before, take $s = (\lambda-\epsilon)\sqrt{n}$ in (4.5). From part (iii), since $\lambda - \epsilon > \frac{e}{\sqrt{2}}$, we have $\lim_{n\to\infty} n\mathbb{P}[M_n \geq (\lambda-\epsilon)\sqrt{n}] = 0$ so m can be selected sufficiently large so that $n\mathbb{P}[M_n \geq s_n] \leq \epsilon\sqrt{n}$ for all $n \geq m$. Part (iv) now follows from (4.5). □

4.2. Yule-Harding Model. For the Yule-Harding model we can derive a recursion for the term $\sum_{t\in RB(s)} \mathbb{P}_s[t]^2$ that occurs in Lemma 4.1.

THEOREM 4.4. *For $s \geq 2$, let $N_s := \sum_{t\in RB(s)} \mathbb{P}_s[t]^2$.*

(i) *N_s satisfies the recursion*

$$N_s = \frac{2}{(s-1)^2}\sum_{r=1}^{s-1}\binom{s}{r}^{-1} N_r N_{s-r} \qquad \text{where } s \geq 2,\ N_1 = 1\,. \tag{4.6}$$

(ii) *Let $a_s = s!N_s$. Then, for $s \geq 2$, a_s satisfies the recursion*

$$a_s = \frac{2}{(s-1)^2}\sum_{r=1}^{s-1} a_r a_{s-r}\,. \tag{4.7}$$

(iii) *Let* $y(x) = \sum_{s=1}^{\infty} a_s x^s$. *Then* $y = y(x)$ *satisfies the differential equation,*

$$x^2 \frac{d^2y}{dx^2} - x\frac{dy}{dx} + y = 2y^2, \qquad y(0) = 0, y'(0) = 1 . \tag{4.8}$$

Proof. We first derive recursion (4.6), then show that $y(x)$ satisfies (4.8). Let t be a phylogeny on $[s]$. Let t_1 and t_2 be the two rooted subtrees of t obtained by deleting the root of t, where we may assume that t_1 has at most half the leaves of T. A fundamental recursion for the Yule-Harding model (Equation 5.3 from [**10**]) states that, for any value of s,

$$\mathbb{P}_s[t] = \frac{2}{(s-1)}\binom{s}{r}^{-1} \mathbb{P}_r[t_1]\mathbb{P}_{s-r}[t_2] .$$

Consequently, by squaring this last equation we see that, for s odd, N_s can be expressed as

$$N_s = \frac{4}{(s-1)^2} \sum_{r<\frac{s}{2}} \binom{s}{r}^{-2} \sum_{t_1,t_2} \mathbb{P}_r[t_1]^2 \mathbb{P}_{s-r}[t_2]^2 ,$$

where $L(t_1), L(t_2)$ form a partition of $[s]$. Thus,

$$\begin{aligned} N_s &= \frac{4}{(s-1)^2} \sum_{r<\frac{s}{2}} \binom{s}{r}^{-1} \sum_{t_1 \in RB(r)} \mathbb{P}_r[t_1]^2 \sum_{t_2 \in RB(s-r)} \mathbb{P}_{s-r}[t_2]^2 \\ &= \frac{4}{(s-1)^2} \sum_{r<\frac{s}{2}} \binom{s}{r}^{-1} N_r N_{s-r} , \end{aligned}$$

which establishes (4.6) when s is odd. For s even the additional term for $r = \frac{s}{2}$ is

$$\frac{1}{2}\frac{4}{(s-1)^2} \sum_{t_1,t_2} \binom{s}{s/2}^{-2} \mathbb{P}_{s/2}[t_1]^2 \mathbb{P}_{s/2}[t_2]^2 ,$$

where t_1, t_2 each have $\frac{s}{2}$ leaves for which the leaf sets form a partition of $[s]$. So, for s even, N_s satisfies the recursion

$$N_s = \frac{4}{(s-1)^2} \sum_{r<s/2} \frac{N_r N_{s-r}}{\binom{s}{r}} + \frac{2}{(s-1)^2}\frac{1}{\binom{s}{s/2}} N_{s/2}^2 .$$

thereby justifying Equation (4.6) in this case also. This completes the proof of part (i).

Part (ii) follows directly from part (i).

For part (iii) note that, for $y(x) = \sum_{s=1}^{\infty} a_s x^s$ we have

$$\sum_{s=1}^{\infty} s a_s x^s = xy'(x) ; \quad \sum_{s=1}^{\infty} s^2 a_s x^s = x^2 y''(x) + xy'(x) . \tag{4.9}$$

Multiplying both sides of (4.7) by $\frac{(s-1)^2}{2} x^s$, then summing over s, gives

$$\frac{1}{2}\sum_{s=1}^{\infty} s^2 a_s x^s - \sum_{s=1}^{\infty} s a_s x^s + \frac{1}{2}\sum_{s=1}^{\infty} a_s x^s = \sum_{s=1}^{\infty} \left[\sum_{r=1}^{s-1} a_r a_{s-r}\right] x^s .$$

Recognising the right-hand side as $y(x)^2$ and using the relationships of (4.9) gives

$$\frac{1}{2}x^2y''(x) - \frac{1}{2}xy'(x) + \frac{1}{2}y(x) = y(x)^2 .$$

Multiplying both sides by two, and setting appropriate initial conditions, completes the proof of part (iii). □

5. Concluding Remarks

It would be instructive to develop an analogue of Theorem 4.3 for the Yule-Harding model by determining the asymptotic behaviour of N_s (perhaps by exploiting Theorem 4.4).

A further interesting problem would be to determine analytical lower bounds (or some power law) on the expected size of $M(T, T')$ under the Uniform or Yule-Harding models.

Results on the expected size of the MAST for three or more random trees may also be of interest.

References

[1] D. Aldous, Probability distributions on cladograms, in (D. Aldous and R. Permantle) **76**, *Random structures*, Springer-Verlag, 1996, 1–18.

[2] A. Amir and D. Keselman, Maximum agreement subtree in a set of evolutionary trees: metrics and efficient algorithms, *SIAM Journal on Computing* **26** (1997), 1656–1669.

[3] D. Bryant, *Building Trees, Hunting for Trees, and Comparing Trees*, Doctoral dissertation, University of Canterbury, Department of Mathematics, 1997.

[4] R. Durret, *Probability: Theory and examples*, ch. 6, 370–371, Duxbury Press, 1996.

[5] W. Feller, *An Introduction to Probability Theory and its Applications*, ch. 11, p. 54, John Wiley and Sons, Inc, 1968.

[6] C. R. Finden and A. D. Gordon, Obtaining common pruned trees, *Journal of Classification* **2** (1985), 255–276.

[7] W. Goddard and K. Grzegorz, The minimum size of agreement subtrees of two binary trees, *Congressus Numerantium*, **97** (1993), 131–136.

[8] W. Goddard, E. Kubicka, and G. Kubicki and F. R. McMorris, The agreement metric for labelled binary trees, *Mathematical Biosciences* **123** (1994), 215–226.

[9] A. Gordon, On the assessment and comparison of classifications, in (R. Tomassine, ed.), *Analyse de Données et Informatique*, Le Chesnay, INRIA, France, 1980, 149–160.

[10] E. F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, *Advances in Applied Probability* **3** (1971), 44–77.

[11] E. Kubicka, G. Kubicki, and F. R. McMorris, On agreement subtrees of two binary trees, *Congressus Numerantium* **88**, 1992, 217–224.

[12] J. B. Losos and F. R. Adler, Stumped by trees? A generalized null model for patterns of organismal diversity, *The American Naturalist* **145** (1995), 329–342.

[13] M. Steel and T. Warnow, Kaikoura tree theorems: computing the maximum agreement subtree, *Information Processing Letters* **48** (1993), 77–82.

[14] M. Steel and A. McKenzie, Properties of phylogenetic trees generated by Yule-type speciation models, *Mathematical Biosciences* **170** (2001), 91–112.

[15] D. Swofford, *PAUP*. Phylogenetic Analysis using Parsimony (*and Other Methods), version 4*, Sinauer Associates, Sunderland, Massachusetts, 1998.

McGill Centre for Bioinformatics, McGill University, Montréal, Québec, Canada

E-mail address: bryant@math.mcgill.ca

LIRMM, 34392 Montpellier, France

E-mail address: andy@lirmm.fr

Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

E-mail address: m.steel@math.canterbury.ac.nz