Mathematical Biosciences xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Does random tree puzzle produce Yule–Harding trees in the many-taxon limit?

Sha Zhu, Mike Steel *

Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

ARTICLE INFO

Article history:
Received 21 August 2012
Received in revised form 2 February 2013
Accepted 8 February 2013
Available online xxxx

Keywords:
Phylogenetic tree
Tree-puzzle
Polyá urn
Centroid vertex

ABSTRACT

It has been suggested that a random tree puzzle (RTP) process leads to a Yule–Harding (YH) distribution, when the number of taxa becomes large. In this study, we formalize this conjecture, and we prove that the two tree distributions converge for two particular properties, which suggests that the conjecture may be true. However, we present statistical evidence that, while the two distributions are close, the RTP appears to converge on a different distribution than does the YH. By way of contrast, in the concluding section we show that the maximum parsimony method applied to random two-state data leads a very different (PDA, or uniform) distribution on trees.

1. Introduction

The Maximum likelihood (ML) approach [4,6,5] is generally considered to be a reliable way of estimating phylogenies from DNA sequences. However, ML is not always feasible for large numbers of species, because of the intensive computation required. Methods that use 'four point subsets' [3] reduce the complexity of the problem, and have assisted numerous studies [2,15,20,21].

The four points subtree is known as the quartet tree. *Quartet puzzling* (QP) [21] is an algorithm to infer a tree on n taxa by using the quartet trees derived from DNA sequences. It firstly computes the likelihood of all $\binom{n}{4}$ quartets. As there are three possible topologies for any four taxa, the quartet tree which returns the greatest ML value is used (any ties are broken uniformly at random). At the puzzling step, the order of inserting new leaf nodes is randomized. A seed tree is built from the first four elements of the ordered leaf node sequence. From this point on, leaves are attached sequentially by the following procedure: when a new leaf x is to be attached to the existing tree T, quartet trees are built from quartets formed from x and all subsets of size three that are chosen from the existing leaf set. If the ML quartet tree of $\{i,j,k,x\}$ is $ij|kx$, then weight 1 is added to the edges on the path in T connecting the two leaves i and j. This process is repeated for all such quartet trees, and x is then attached to the edge which has the minimal weight. An example is given in Fig. 1.

Since the order of adding leaves is randomized, this can lead to variation in the resulting tree topologies, and so a consensus tree of numerous replicates is used as the output tree. The program *Tree puzzle* (TP) [16] is a parallel version of QP, which performs independent puzzling steps simultaneously.

The trees generated by either the QP or TP process depend on the biological sequences we have for the taxa. To investigate how the TP process behaves on randomized quartets, Vinh et al. [22] performed a simulation study on a so-called *random tree puzzle* (RTP) process. This assumes that no prior molecular information is given. Therefore, for the same quartet set, all three tree topologies are equally likely. The authors compare the empirical probabilities of tree topologies against the theoretical probabilities from the *proportional to distinguishable arrangement* (PDA) model and the *Yule–Harding* (YH) model. Table 1 from [22] reveals that the RTP's empirical probabilities are very close to the YH theoretical probabilities (indeed, there are two cases where these probabilities are identical). As it seems that the differences between the empirical and theoretical probabilities decrease as the number of taxa increases, Vinh et al. [22] suggest that the RTP process converges to the YH process as n (the number of taxa) grows. The authors provided further evidence for their conjecture by comparing some properties of RTP trees with YH trees. Recall that a *cherry* in a tree is a pair of leaves that are adjacent to the same vertex. Then Vinh et al. [22] found that the mean and variance of the number of cherries were similar under the RTP simulation and the theoretical value under the YH process [13].

Although Vinh et al. [22] provided evidence to suggest the two distributions appear to become very similar as n grows, they did not provide a formal statement or proof of their claim that the two distributions converge. In this project, we investigate the RTP process further using mathematical and statistical methods. Our results demonstrate that certain properties of the trees that are near the 'periphery' of the tree (i.e. near the leaves) converge

* Corresponding author. Tel.: +64 21329705.
E-mail addresses: sha.joe.zhu@gmail.com (S. Zhu), mike.steel@canterbury.ac.nz (M. Steel).

0025-5564/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2013.02.003

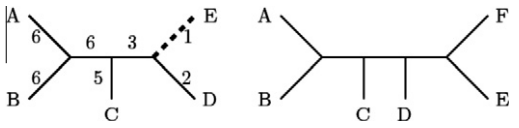

Fig. 1. Suppose leaf F is about to be attached to the five-taxon tree on the left, and the ML trees of $\{i,j,k,F\}$ are $AB|CF, AC|EF, BC|DF, AC|DF, AB|DF, AD|EF, AB|EF, BC|EF, BD|EF$, and $CE|DF$. The external edge leading to E returns the minimal weight, so F is attached to this edge, leading to the six-taxon tree shown on the right.

under the two distributions; however the 'deep' structure of the trees (how the tree is broken up around its centroid) appears to retain a trace that distinguishes the two models as the trees become large.

A key technique we employ is to describe aspects of the tree as it evolves (by the random addition of taxa under either a YH or RTP process) via Polyá urn type models. Roughly speaking, these models describe the distribution of balls of various colours in an urn under a sequential sampling scheme whereby, at each step, a ball is drawn at random and then replaced by certain number of balls of different colours (dependent on the colour of the ball drawn). There is a well-developed asymptotic theory for such models in the probability literature (an excellent recent survey can be found in [12]. In [13] this theory (for balls of two colours) is used to show how the distribution of cherries in YH trees is asymptotically normal with an easily computed mean and variance. Here we use this asymptotic theory in two further ways to study the RTP process – firstly, by a similar argument to [13] to study the distribution of cherries, again using balls of two colours, and also (using balls of three colours) to consider the distribution of leaves about the three subtrees incident with a fixed vertex of a tree under a YH process of leaf addition.

Computing the distribution of trees exactly under the RTP process is possible for small trees, but becomes problematic as the number of leaves grows. This is in contrast to the PDA or YH models, where simple explicit formulae are on hand to compute the probability of any rooted binary tree (see e.g. [17]. One reason for the additional complexity of computing the tree distribution under the RTP process is the minimization step that arises in selecting the edges of the tree to which the next leaf is to be attached.

In summary, an outline of our results is as follows. Firstly we formalize two versions of the conjecture, and establish the weaker form, which shows that the 'peripheral' structure of the tree behaves exactly like the YH model as the number of leaves tends to infinity. We then investigate the stronger conjecture, where the 'deep' structure of a tree becomes important. We provide statistical evidence that a model (we call RTP′) that is a hybrid between RTP and YH behaves differently from YH in the limit as the number of leaves grows, which suggests that the stronger conjecture is likely to be false.

2. Formalized conjectures

Given two discrete probability distributions p and q on a finite set Y, the *total variational distance* between p and q is defined as:

$$d_{\mathrm{VAR}}(p,q) = \max_{A \subseteq Y} |\mathbb{P}_p(A) - \mathbb{P}_q(A)|,$$

where $\mathbb{P}_p(A) = \sum_{y\in A} p(y)$ and $\mathbb{P}_q(A) = \sum_{y\in A} q(y)$ are the probabilities of event A under the distributions p and q respectively. Thus $d_{\mathrm{VAR}}(p,q)$ is the largest possible probability difference of any event under the distributions p and q. A well-known and elementary result is that $d_{\mathrm{VAR}}(p,q) = \frac{1}{2}\sum_{y\in Y}|p(y) - q(y)|$, and thus the two distributions are the same if and only if $d_{\mathrm{VAR}}(p,q) = 0$.

A tree with the leaf set $X_n = \{1,2,\ldots,n\}$ is called an X_n-*tree*. In the rest of this article, all X_n-trees referred to are binary trees, where the interior nodes have degrees of three. We use T_n to denote a labeled X_n-tree topology, and t_n to denote an unlabeled X_n-tree shape. [22] suggest that when the number of taxa (n) becomes large, RTP converges to the YH distribution. In this study, we consider the total variational distance between the two probability distributions on X_n-trees generated by the RTP and the YH process, and we formalize the conjecture from [22]. This formalization states that the variational distance between the two tree distributions converges to zero as the number of taxa added grows. We first note that it makes no difference to the truth of this conjecture whether the trees are labeled or unlabeled.

Lemma 1. *Let $\mathcal{T}(n)$ and $\mathcal{S}(n)$ be the set of labeled and unlabeled X_n-trees respectively. For $T_n \in \mathcal{T}(n)$, and $t_n \in \mathcal{S}(n)$, let $\Delta_n := \sum_{T_n\in\mathcal{T}(n)} |\mathbb{P}_{YH}(T_n) - \mathbb{P}_{RTP}(T_n)|$ and $\delta_n := \sum_{t_n\in\mathcal{S}(n)} |\mathbb{P}_{YH}(t_n) - \mathbb{P}_{RTP}(t_n)|$. Then, $\Delta_n = \delta_n$, and in particular $\lim_{n\to\infty}\Delta_n = 0 \Longleftrightarrow \lim_{n\to\infty}\delta_n = 0$, as $n \to \infty$.*

Proof. Let $v(t_n)$ be the number of X_n-trees T_n that have the shape t_n. Then, for $* \in \{\mathrm{YH}, \mathrm{RTP}\}$, $\mathbb{P}_*(T_n) = \frac{\mathbb{P}_*(t_n)}{v(t_n)}$, we have:

$$\begin{aligned}
\Delta_n &= \sum_{T_n\in\mathcal{T}(n)} |\mathbb{P}_{\mathrm{YH}}(T_n) - \mathbb{P}_{\mathrm{RTP}}(T_n)| \\
&= \sum_{t_n\in\mathcal{S}(n)} \sum_{\substack{T_n\in\mathcal{T}(n) \\ T_n \text{ has shape } t_n}} |\mathbb{P}_{\mathrm{YH}}(T_n) - \mathbb{P}_{\mathrm{RTP}}(T_n)| \\
&= \sum_{t_n\in\mathcal{S}(n)} v(t_n)\left|\frac{\mathbb{P}_{\mathrm{YH}}(T_n)}{v(t_n)} - \frac{\mathbb{P}_{\mathrm{RTP}}(T_n)}{v(t_n)}\right| = \sum_{t_n\in\mathcal{S}(n)} |\mathbb{P}_{\mathrm{YH}}(t_n) - \mathbb{P}_{\mathrm{RTP}}(t_n)| \\
&= \delta_n. \qquad \square
\end{aligned}$$

Thus, we formalize the conjecture from Vinh et al. [22] as follows:

Conjecture (strong version)
With $\Delta_n = \delta_n$ defined as above, $\lim_{n\to\infty}\Delta_n = 0$.

Note that, in the YH process, new leaves are only ever attached to pendant edges, and each pendant edge is selected with equal probability. We say that such leaves are attached to *uniformly selected pendant edges*. By contrast, the RTP process can attach new leaves to any edge, although RTP has an increasingly strong preference to attach leaves to pendant edges as the tree grows [22]. These authors also suggested that as the tree grows, the number of cherries of a RTP tree follows the same limiting distribution as the number of cherries of a YH tree, which is normally distributed. We summarize these two claims as follows:

Conjecture (weak version)

1. Let $\mathcal{E}_m$ be the event that *all* leaf attachments under the RTP beyond the first m leaves, are to uniformly selected pendant edges. Then $\mathbb{P}(\mathcal{E}_m) \to 1$, as m tends to infinity.
2. The distribution of cherries converges to the same (asymptotic) normal distribution as the YH model.

In our paper, we prove the two parts of the weak conjecture, and present statistical evidence that the strong conjecture is not true.

3. RTP is similar to YH when *n* is large

To verify Part 1 of the weak conjecture, we need to establish that the probability that a new leaf attaches to a pendant edge converges to 1 sufficiently quickly as the number of leaves increases. This requires that the pendant edges carry less weight than the inte-

Please cite this article in press as: S. Zhu, M. Steel, Does random tree puzzle produce Yule–Harding trees in the many-taxon limit?, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.02.003

rior edges. In addition, when the new leaf is added, all pendant edges must be equally likely to be chosen. Thus we must check the edge weight distribution during the puzzling step of the RTP process.

3.1. Distribution of edge weights

Let E_n^{p} denote the set of pendant edges of the current X_n-tree T_n and let E_n^{I} be the set of interior edges. For any edge e of T_n, we let $W(e)$ denote the random variable edge weight during the quartet puzzling step. Suppose edge e has k leaves of T_n on one side and $n-k$ leaves of T_n on the other side. The following result is established in the Appendix.

Lemma 2. *$W(e)$ is a binomial random variable with the parameters $\frac{k(n-k)(n-2)}{2}$ as the number of trials and $\frac{2}{3}$ as the probability of success on each trial.*

The parameter k takes the value 1 or $n-1$ for a pendant edge; for an interior edge, k lies between 2 an $n-2$. Next, we show that for any fixed pendant and interior edge, the probability that the interior edge has lower weight converges to zero exponentially fast with increasing n. More precisely, for any $e'' \in E_n^{\mathrm{p}}$ and any $e' \in E_n^{\mathrm{I}}$, we establish the following result in the Appendix.

$$\mathbb{P}(W_n(e'') \geqslant W_n(e')) \leqslant 2\exp\left(-\frac{1}{144}n\right). \tag{1}$$

This result is for a fixed pair of pendant and interior edges, but it easily implies that the probability that the smallest weight in the tree is on a pendant rather than an interior edge converges quickly to 1 with increasing n. This is formalized in the following inequality, also proved in the Appendix:

$$\mathbb{P}\left(\min_{e\in E_n^{\mathrm{p}}}\{W_n(e'')\} \leqslant \min_{e'\in E_n^{\mathrm{I}}}\{W_n(e')\}\right) \geqslant 1 - 2n^2\exp\left(-\frac{1}{144}n\right). \tag{2}$$

Thus a new leaf is almost certain to be added to pendant edges; moreover, as noted above, each pendant edge has equal probability of being attached to.

3.2. New leaves attach rarely to interior edges

Theorem 1. *Suppose $T_m \in \mathcal{T}(m)$, let $\mathcal{E}_m$ be the event that all leaf attachments under RTP beyond T_m are to uniformly selected pendant edges. Then, for constants $a, b > 0$:*

$$\mathbb{P}(\mathcal{E}_m) \geqslant 1 - ae^{-bm}.$$

Proof. Let B_k be the event that $(k+1)$−st leaf is not attached to any leaf edge of T_k. Then we have $1-\mathbb{P}(\mathcal{E}_m) = \mathbb{P}(\bigcup_{k=m}^{\infty} B_k)$. By Boole's inequality, we have $\mathbb{P}(\bigcup_{k=m}^{\infty} B_k) \leqslant \sum_{k=m}^{\infty}\mathbb{P}(B_k)$. By Inequality (2), $\mathbb{P}(B_k) \leqslant 2k^2\exp\left(-\frac{1}{144}k\right)$. We now use the following general inequality, the proof of which is given in the Appendix. If $Q_m = \sum_{k=m}^{\infty} k^2\exp(-ck)$, where $c \geqslant \frac{4\log k}{k}$ and $k > 1$, then for $m \geqslant m_0$:

$$Q_m \leqslant \frac{\exp(-cm_0/2)}{1-\exp(-c/2)}. \tag{3}$$

Thus,

$$\begin{aligned} 1-\mathbb{P}(\mathcal{E}_m) &\leqslant \sum_{k=m}^{\infty} 2k^2\exp\left(-\frac{1}{144}k\right) \\ &\leqslant \frac{2}{1-\exp\left(-\frac{1}{288}\right)}\exp\left(-\frac{1}{288}m\right). \end{aligned}$$

Rearranging this inequality establishes the inequality in the theorem. The uniformity follows by Lemma 2. □

3.3. The mean and variance of the number of cherries in the RTP tree

Table 3 of Vinh et al. [22] reveals that the mean and variance of the number of cherries on trees generated under the RTP process and under YH process are similar. In order to provide a formal proof that they converge to the same limiting distribution, we need to introduce the *extended Polyá urn model* (EPU).

3.3.1. Extended Polyá urn model

Consider the following extended Polyá urn (EPU) model: at time $t=0$, there are b blue balls and r red balls in an urn, where $b \geqslant 0$ and $r \geqslant 0$. At each discrete time step, one ball is picked at random from the urn. If the ball is blue, c additional blue balls and d red balls will be placed; if the picked ball is red, e additional blue balls and f red balls will be placed. The values c, d, e, f can also take negative values, in which case, instead of placing new balls in the urn, the number of balls of the appropriate colour will be withdrawn. We use b_n to denote the number of blue balls after the nth draw, and S_n is the total number of balls. The following matrix describes this process:

$$A = \begin{bmatrix} c & d \\ e & f \end{bmatrix}.$$

We require that A has positive and equal row sums, as well as one real positive principal eigenvalue λ. Let $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be the normalized eigenvector associated with λ. Then, under these conditions, a classic result [12,18] states that, as $n \to \infty$, $\frac{b_n - \lambda v_1 n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2)$ [12,1], where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. Crucially, the initial values of b and r do not play any significant roles in this limiting normal distribution (or of its mean and variance).

3.3.2. EPU and attaching new edges only to pendant edges

We relate the Yule process to the EPU model as follows: consider the set of cherry edges as a collection of blue balls, and the non-cherry edges as a collection of red balls. When a new edge is attached to a pendant edge, if it is attached to a cherry edge, the number of cherry edges remain the same, but the number of non-cherry edges increases by one. If a new edge is added to a non-cherry edge, then the non-cherry edge becomes a cherry edge, and the new edge is also a cherry edge. Thus, the generating matrix is:

$$A = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}.$$

Notice that A has row sum equal to 1 and A has one real positive eigenvalue λ, as required.

Let C_n be the number of cherries in a YH tree. Then as n tends to infinity,

$$Z_n := (C_n - n/3)/\sqrt{2n/45}$$

converges in distribution to a standard normal distribution (i.e. $Z_n \xrightarrow{\mathcal{D}} N(0,1)$), by Corollary 3 of [13]. We now show that the same holds for the distribution of cherries in an RTP tree.

Theorem 2. *Let C_n^* be the number of cherries in an RTP tree, and let $Z_n^* = (C_n^* - n/3)/\sqrt{2n/45}$. Then $Z_n^* \xrightarrow{\mathcal{D}} N(0,1)$.*

Proof. We need to show that for any $\epsilon > 0$, and for all sufficiently large value of n and all positive real x,

$$|\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z < x)| \leqslant \epsilon, \tag{4}$$

where Z is a standard normal random variable.

Please cite this article in press as: S. Zhu, M. Steel, Does random tree puzzle produce Yule–Harding trees in the many-taxon limit?, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.02.003

As before, let $\mathcal{E}_m$ be the event that after m leaves have been attached to the starting tree by RTP, all further additions are to pendant edges, and let $\mathcal{E}_m^c$ be the complement of $\mathcal{E}_m$. For $n > m$, by the law of total probability, we have:

$$\mathbb{P}(Z_n^* < x) = \mathbb{P}(Z_n^* < x|\mathcal{E}_m)\mathbb{P}(\mathcal{E}_m) + \mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c). \quad (5)$$

If we now subtract $\mathbb{P}(Z_n^* < x|\mathcal{E}_m)$ from both side of Eq. (5), we obtain:

$$\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n^* < x|\mathcal{E}_m) = \mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1) + \mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c). \quad (6)$$

By the triangle inequality ($|a+b| \leqslant |a| + |b|$) we have:

$$|\mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1) + \mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c)| \leqslant |\mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1)| + |\mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c)|. \quad (7)$$

Combining Eq. (6) and Inequality (7) gives the following:

$$|\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n^* < x|\mathcal{E}_m)| \leqslant |\mathbb{P}(Z_n^* < x|\mathcal{E}_m)(\mathbb{P}(\mathcal{E}_m) - 1)| + |\mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)\mathbb{P}(\mathcal{E}_m^c)|, \leqslant |\mathbb{P}(Z_n^* < x|\mathcal{E}_m)||(\mathbb{P}(\mathcal{E}_m) - 1)| + |\mathbb{P}(Z_n^* < x|\mathcal{E}_m^c)||\mathbb{P}(\mathcal{E}_m^c)|. \quad (8)$$

Theorem 1 tells us that $\mathbb{P}(\mathcal{E}_m) \geqslant 1 - ae^{-bm}$, which tends to 1 as m grows. Now, since $\mathbb{P}(\mathcal{E}_m^c) \to 0$ as m tends to infinity, we can select a sufficiently large value of m that $\mathbb{P}(\mathcal{E}_m^c) \leqslant \epsilon/4$ and $\mathbb{P}(\mathcal{E}_m) \geqslant 1 - \epsilon/4$. Thus, $\mathbb{P}(\mathcal{E}_m) - 1 \geqslant -\epsilon/4$, and $|\mathbb{P}(\mathcal{E}_m) - 1| \leqslant \epsilon/4$. Since $0 \leqslant \mathbb{P}(Z_n^* < x|\mathcal{E}_m), \mathbb{P}(Z_n^* < x|\mathcal{E}_m^c) \leqslant 1$, Inequality (8) gives:

$$|\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n^* < x|\mathcal{E}_m)| \leqslant \epsilon/4 + \epsilon/4 = \epsilon/2, \quad (9)$$

for all sufficiently large m, and all $n \geqslant m$ and $x > 0$.

Now we consider the sequence of Z_n^* conditional on $\mathcal{E}_m$. By conditioning on this event all the new leaves are to uniformly selected pendant edges. Because the EPU argument that established the convergence of the sequence Z_n (the normalization of the number of cherries in a YH tree) does not depend on the initial number of cherries for any $\epsilon > 0$, and every m, there exists an integer n_0 so that for all $n \geqslant n_0$, and $x > 0$:

$$|\mathbb{P}(Z_n^* < x|\mathcal{E}_m) - \mathbb{P}(Z_n < x)| \leqslant \epsilon/2. \quad (10)$$

Then, by the triangle inequality ($|a+b| \leqslant |a| + |b|$), if we add Inequalities (9) and (10), we have

$$|\mathbb{P}(Z_n^* < x) - \mathbb{P}(Z_n < x)| \leqslant \epsilon$$

and since Z_n converges in distribution to a standard normal, this establishes (4). □

Theorem 2 shows that the number of cherries on the RTP trees has a limiting normal distribution with the same asymptotic mean and variance as for the YH distribution.

We have also shown that, from some point onward, new leaves will always be added to pendant edges, which verifies the weak conjecture. While these two results may be regarded as providing some weak evidence in favour of the strong conjecture, they do not constitute any formal justification of it. In the next section, we will provide an analysis which suggests that the variational distance between the two distributions remains bounded away from zero as n grows, and this makes these two process distinct in the limit.

4. Is RTP the same as YH?

Consider the following scenario where we perform the YH process on some starting tree with more than three leaves, where v is one of the interior nodes. At node v, the graph is divided into three subtrees (see Fig. 2). We let $L_i, (i = 1, 2, 3)$ denote the leaf sets of these subtrees, and let $l_i = |L_i|, (i = 1, 2, 3)$ denote the number of leaves in the sets. We normalize the l_i values by the total number of leaves n. Clearly, the sequence of l_i/n values change, as new leaves are gradually added to the whole tree.

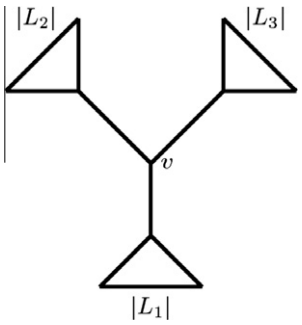

Fig. 2. Centroid of a tree.

4.1. Polyá urns and the centroid of a tree

Adding new leaves onto the tree under the YH process ensures that each new leaf is always added into one of the leaf sets $L_i, (i = 1, 2, 3)$. The probability that l_i increases by one is the relative proportion of the number of leaves of the subtree in relation to the number of leaves in the full tree. This is similar to the Polyá urn problem [10] involving balls of three different colours.

Suppose that one ball is picked randomly at each step, and replaced along with another ball of the same colour into the urn. Let F_n^i be the relative frequency of the ith colour ball when n balls are present, and $\mathbf{F}_n = (F_n^1, F_n^2, F_n^3)$. Then $\mathbf{F}_n$ converges (as $n \to \infty$) to a Dirichlet distribution [11] with the parameter vector $\mathbf{F}_{n_0}$, where n_0 is the total initial number of balls. Different initial values in the urn produce different distributions when n balls are present in the urn, and this difference in distributions does not converge to zero as n grows. This result suggests that the YH process on different initial X-trees may well lead to different distributions of the resulting trees. However, if the final tree shape is the only information we are given, then it will be impossible to identify the position of the original vertex v in the final tree with certainty. Thus the frequencies $\mathbf{F}_n$ cannot be clearly measured from the final tree alone. However, we can partly ameliorate this problem by considering a particular vertex that we can easily identify in the final tree, namely its centroid [8,14].

Definition. A vertex v of a tree $T = (V, E)$ is a *centroid* if each component of the disconnected graph $T \setminus v$ has, at most $(1/2)|V|$ vertices.

A well known property of centroids states that a tree has either a single centroid or two adjacent centroids, in which case $|V|$ is even [9]. To keep the problem simple, we only consider trees with a single centroid. However, because T is a binary tree, $|V|$ is always even, and so this does not guarantee a unique centroid. Fortunately, the following lemma shows that a binary tree with odd number of leaves always has a unique centroid.

Lemma 3. *Let T be an unrooted binary X_n-tree. Then:*

1. *A vertex v of T is a centroid of T if and only if v satisfies $l_1, l_2, l_3 \leqslant \frac{n}{2}$, where l_i are the number of leaves of the three subtrees of $T \setminus v$.*
2. *If n is odd, then T has a unique centroid.*

Proof.

(1) Suppose that v is an interior vertex of T. Consider the vertex sets V_1, V_2 and V_3 of the connected components of $T \setminus v$. Let l_i be the number of leaves in V_i. Considering the rooted binary tree on V_i, we have:

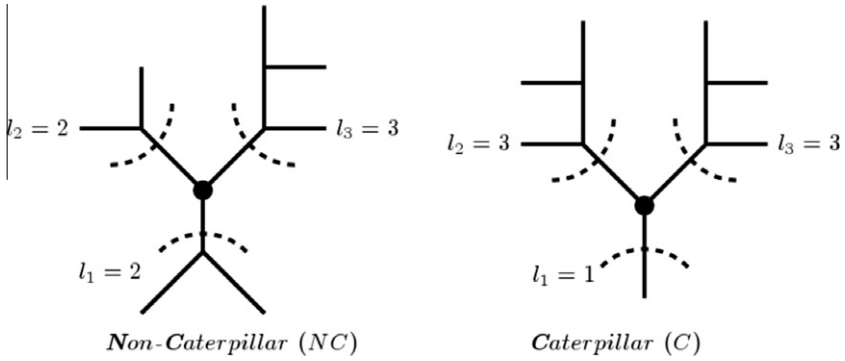

Fig. 3. The two tree shapes for binary trees on seven leaves.

$$|V_i| = 2l_i - 1. \tag{11}$$

Also, since T is an unrooted binary tree, we have:

$$|V| = 2n - 2. \tag{12}$$

Thus, $|V_i| \leqslant \frac{1}{2}|V|$ if and only if $2l_i - 1 \leqslant \frac{1}{2}(2n-2)$ and this holds precisely if $l_i \leqslant n/2$. Thus, the condition for v to be a centroid (namely that $|V_i| \leqslant \frac{1}{2}|V|$ for $i = 1,2,3$) is precisely the same as that stated in the lemma.

(2) Suppose v is a centroid of T. At v, we let L_i, $(i = 1,2,3)$ denote the leaf set of the subtrees T_i and let l_i denote the size of these leaf sets, ordered so that $l_j \leqslant l_3 \leqslant \frac{|X|}{2}$, $(j = 1,2)$. Since n is odd, we have $l_3 < \frac{n}{2}$.
Suppose another centroid d exists. We use L'_i to denote the complement of L_i. Then there is a subtree H of T rooted at d, with leaf set L_H, where $L_H \supseteq G'$, and $G' \in \{L'_1, L'_2, L'_3\}$. Since $l_j \leqslant l_3 < \frac{n}{2}$, where $j \in \{1,2\}$, we then have $|L_H| \geqslant |G'| > \frac{n}{2}$. Therefore, d cannot be a centroid. □

We now relate the centroid back to the Polyá urn problem. First notice that tree shapes only start to differentiate when there are more than five leaves. Therefore, in the following scenario, we perform the YH process from initial trees with seven leaves (we start with trees with seven leaves, rather than six, as we wish to restrict attention to trees with an odd number of leaves, and which therefore have a unique centroid). Suppose that a tree X is either the non-caterpillar (NC) or caterpillar (C) tree shown in Fig. 3. We will use X as the initial tree to construct some tree t_n. At the centroid of t_n when $n = 7$ the sequences of l_i/n are $(2/7, 2/7, 3/7)$ and $(1/7, 3/7, 3/7)$ for $t_7 = NC$ and $t_7 = C$ respectively. Now, let us only consider the number of leaves l_1 in the smallest subtree of t_n for all odd values of $n \geqslant 7$ (henceforth all values of n in this section are odd to guarantee a unique centroids, and limits as n tends to infinity are also over just the odd values of n). We define the ratio of l_1 and of number of leaves n as $\pi_n^X = \frac{l_1}{n}$. For $\gamma \in (0,1)$, let Π^X be the limiting probability of the event $\pi_n^X \geqslant \gamma$. In other words, $\Pi^X = \lim_{n\to\infty} \mathbb{P}(\pi_n^X \geqslant \gamma)$. To test the null hypothesis that $\Pi^{NC} = \Pi^C$, we investigate the ratio π_n^X under the YH process starting with a tree on 7 leaves with shape $X \in \{C, NC\}$. An additional 2000 leaves are attached to the starting trees of shape NC and C under the YH process with 1000 replicates in each case[1]. Using the initial tree of shape NC or C, we found that the probability that π_n^X is greater than $\gamma = 0.19$ does not appear to be converging for the two choices of X (NC or C) (see Fig. 4). Fig. 4 indicates the 95% confidence interval of proportions of the event for which $\pi_n^X \geqslant 0.19$, which suggests the following strict inequality:

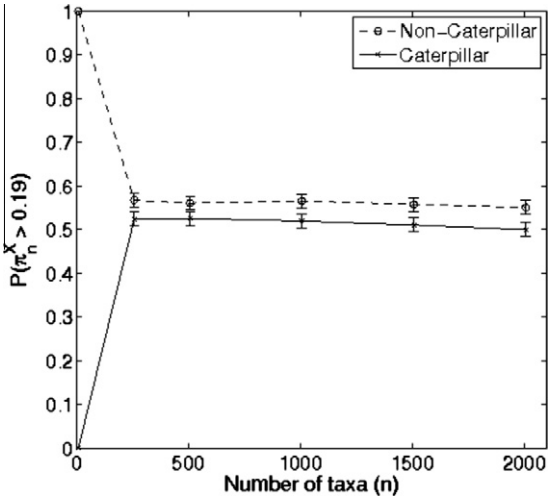

Fig. 4. Empirical probabilities and the 95% confidence interval proportion of the event that $\pi_n^X \geqslant 0.19$. The dashed line is for the initial tree of the non-caterpillar seven-taxa tree; and solid line is for the caterpillar seven-taxa tree.

$$\Pi^{NC} > \Pi^C. \tag{13}$$

4.2. A modified RTP process

To provide statistical evidence that the RTP and the YH processes are not exactly the same, we define a new process **RTP′**, which is equivalent to the RTP process up to $n = 7$. From this point forward it proceeds according to the YH process. Therefore, the initial probabilities of constructing X_n-trees from NC and C under the RTP′ process are different from the YH process. We use the probabilities of the starting tree NC and C under the RTP process as the probabilities under the RTP′. [22] estimated by simulations that the probabilities for the seven-taxa non-caterpillar tree is 0.4607 under the RTP process and 0.4667 under the YH process, which gives us the following inequality:

$$\mathbb{P}_{YH}(t_7 = NC) - \mathbb{P}_{RTP'}(t_7 = NC) > 0. \tag{14}$$

Theorem 3. *If (13) holds then*

$$\lim_{n\to\infty} d_{\mathrm{VAR}}(\mathbb{P}_{\mathrm{RTP}'}(t_n), \mathbb{P}_{\mathrm{YH}}(t_n)) \neq 0.$$

Proof. Let $\mathcal{S}(n)$ be the set of unlabeled X_n-tree and let:

$$\delta' := \sum_{t_n \in \mathcal{S}(n)} |\mathbb{P}_{\mathrm{YH}}(t_n) - \mathbb{P}_{\mathrm{RTP}'}(t_n)|. \tag{15}$$

[1] The software ('sim_tree') for this analysis is available at http://www.math.canterbury.ac.nz/bio/downloads

Please cite this article in press as: S. Zhu, M. Steel, Does random tree puzzle produce Yule–Harding trees in the many-taxon limit?, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.02.003

For a tree generated by YH or RTP′, consider the event Σ_n that $\pi_n \geqslant \gamma$, where $\pi_n = l_1/n$ is the proportion of leaves of the tree when it has n leaves, that lie in the smallest subtree(s) incident with the centroid. Then:

$$\mathbb{P}_{\mathrm{YH}}(\Sigma_n) = \sum_{X\in\{NC,C\}} \mathbb{P}_{\mathrm{YH}}(\Sigma_n|t_7 = X)\mathbb{P}_{\mathrm{YH}}(t_7 = X), \tag{16}$$

$$\mathbb{P}_{\mathrm{RTP}'}(\Sigma_n) = \sum_{X\in\{NC,C\}} \mathbb{P}_{\mathrm{RTP}'}(\Sigma_n|t_7 = X)\mathbb{P}_{\mathrm{RTP}'}(t_7 = X). \tag{17}$$

If we now subtract Eq. (17) from (16), and substitute $\mathbb{P}_*(t_7 = C)$ in $1 - \mathbb{P}_*(t_7 = NC)$, we have:

$$\mathbb{P}_{\mathrm{YH}}(\Sigma_n) - \mathbb{P}_{\mathrm{RTP}'}(\Sigma_n) = (\mathbb{P}_{\mathrm{YH}}(t_7 = NC) - \mathbb{P}_{\mathrm{RTP}'}(t_7 = NC))(\Pi^{NC} - \Pi^{C}). \tag{18}$$

Thus, if we apply Inequalities (14) and (13) in Eq. (18), we obtain.

$\mathbb{P}_{\mathrm{YH}}(\Sigma_n) - \mathbb{P}_{\mathrm{RTP}'}(\Sigma_n) > 0$. Consequently, $\delta' > 0$ in (15), and so $\lim_{n\to\infty} d_{\mathrm{VAR}}(\mathbb{P}_{\mathrm{RTP}'}(t_n), \mathbb{P}_{\mathrm{YH}}(t_n)) \neq 0$, as claimed. □

It is important to be clear about what we have established: we have not formally shown that RTP does not converge to YH, nor even that RTP′ fails to converge to YH. Rather, we have provided statistical evidence that a certain property of RTP′ holds, and if so, this implies (Theorem 3) that RTP′ does not converge to YH. Then, since RTP′ is a hybrid of YH and RTP, this suggests that RTP does not either.

5. Further discussion and concluding comments

In phylogenetic studies, trees are inferred from DNA sequences using various methods. It is also pertinent to ask what sort of trees these methods would produce, given entirely random data. This is one of the motivations of the study by [22]. In the following discussion, we use an n by k matrix D to denote a sequence of k independent characters on n taxa. Note that all the characters have the same state space S. The term 'random data' can refer to any one of the following three schemes:

(R1) State x is assigned to taxon i in character j by an independent, identically distributed (i.i.d.) process with a probability $p_j(x)$, for $x \in S$.

When the probabilities of state x are the same for all characters (i.e. if $p_j(x) = p(x)$ for all j), we obtain a stronger notion as follows:

(R2) For every entry of the matrix D, D_{ij} is assigned to state x with probability $p(x)$.

If all states are equally likely (i.e. if $p(x) = 1/|S|$), we arrive at an even stronger notion as follows:

(R3) For all entries of D, all states have equal probabilities.

Vinh et al. [22] suggest that random data imply that quartet trees are equally likely and independent to each other, stating:

> In our setting, we assume no phylogenetic information in the data. This is equivalent to the assumption that each of the three topologies for a quartet is equally likely and that the tree topology for each quartet is independent of the other quartets. ...Hence, $3^{\binom{n}{4}}$ possible combinations of quartet trees will serve as input to TP.

For any of the models (R1)–(R3), it certainly is true that random sequence data provide equal support for all three possible topologies of any four taxa. However, this does not necessarily imply that the inferred quartet trees are exactly independent. Rather than persue this question here, we will consider the behaviour of TP under a model in which quartet trees are i.i.d. and uniform, as in [22].

While the RTP process appears to converge close to the YH distribution, it is instructive to note that another tree reconstruction method, *maximum parsimony* (MP), when applied on random data, converges to a quite different distribution on trees. Under model (R3) with two states MP converges to the PDA ('proportional to distinguishable arrangements') model, which selects each unrooted binary tree with equal probability. Let $B(n)$ be the set of unrooted binary trees on the leaf set $\{1, 2, \ldots n\}$. For model (R3) with two states and k independent characters, we use $\mathcal{T}_{MP}(D)$ to denote the MP tree on D (if the MP tree for D is not unique then select one MP tree uniformly at random).

Theorem 4. *Under random model* **(R3)** *with two states:*

1. *The random tree $\mathcal{T}_{MP}(D)$ has a PDA distribution on $B(n)$; i.e.*

$$\mathbb{P}(\mathcal{T}_{MP}(D) = T) = \frac{1}{|B(n)|}.$$

2. *For each fixed n, there is a unique MP tree for D with probability converging to 1 as k grows.*

Proof.

1. Let $w(D,T), T \in B(n)$, denote the parsimony score of T on random data D. By Theorem 7.1 of [19], the number of ways to colour the leaves of a binary tree T with n leaves with using two colours, and so that the resulting colouration has parsimony score of k for T depends only on n and not otherwise on the tree T. Hence, for all $T \in B(n)$, the probability $\mathbb{P}(w(D,T) = l)$, is the same for all binary trees with a given number of leaves. Therefore, each tree has the same probability of being an MP tree for D.
   Let $E_k(T,T')$ be the event that T and T' have exactly the same parsimony score. By the Central Limit Theorem, the probability that the difference in scores is exactly 0 (i.e. $\mathbb{P}(E_k(T,T'))$) tends to zero as k grows.
   Let E be the event that the maximum parsimony tree for D is unique, and let E^c be the complement, namely that there are at least two trees which have the same parisimony score for D. Note that E^c is a subset of the union of the events $E_k(T,T')$ over all T, T' (distinct). Therefore, we have:

$$1 - \mathbb{P}(E) \leqslant \mathbb{P}\left(\bigcup_{T,T'} E_k(T,T')\right) \leqslant \sum_{T,T'} \mathbb{P}\big(E_k(T,T')\big) \to 0,$$

as k grows. Thus, $\mathbb{P}(E) \to 1$, as $k \to \infty$, as required. □

Hence the MP tree on random data with two states converges to the PDA model.

In the PDA model, new leaf nodes are uniformly added onto any edges of the existing tree, whereas the Yule tree selects a pendant edge randomly, and adds a new node onto this pendant edge. During the construction process, PDA, RTP and RTP′ can attach some new leaves onto interior edges. For the PDA process, this has probability of almost 1/2, and it is much less for RTP, as the number of leaves increases. In the case of RTP′, beyond seven leaves, all further leaves are inserted to a pendant edge, just as in the YH model.

In conclusion, we have verified that the RTP process will eventually not add new leaves onto interior edges after some point, which makes the RTP process become more like the YH process. However,

Please cite this article in press as: S. Zhu, M. Steel, Does random tree puzzle produce Yule–Harding trees in the many-taxon limit?, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.02.003

the distance between two distributions appears to remain bounded away from zero even when n tends to infinity, which suggests that they are still two distinct tree construction methods. Nevertheless the RTP process will produce more balanced trees than methods such as maximum parsimony on certain (random) data. It would therefore be interesting to compare the shapes of phylogenetic trees constructed from real biological data using these two tree reconstruction methods to determine if there is a systematic difference in tree shape balance between the two methods on identical data sets. A further, more theoretical project for future work would be to obtain an exact asymptotic analysis of the distribution of the three subtrees about the centroid vertex under the RTP′ process.

Acknowledgments

We thank Marsden Fund for supporting this work. We also thank David Aldous for suggesting we consider the stochastic properties of RTP′ trees relative to their centroids. We also thank the two anonymous reviewers for several helpful suggestions on an earlier version of this manuscript.

Appendix: Technical details

Justification of Lemma 2

Proof. At edge e, suppose that A and B partition X_n, where $n-1 \geqslant k \geqslant 1, |A| = k$ and $|B| = n-k$. Let $\{a,b,c\}$ be a subset of X_n of size three. Suppose that a new leaf x is to be attached to e. Let q be a split of $\{x,a,b,c\}$, and $q = xc|ab, xa|bc, xb|ac$ with equal probabilities. Suppose, a and b are always on one side of e, we consider the following four cases,

$$\begin{cases} \text{case I} & c \in B \text{ and } \{a,b\} \subseteq A; \\ \text{case II} & \{a,b,c\} \subseteq B; \\ \text{case III} & c \in A \text{ and } \{a,b\} \subseteq B; \\ \text{case IV} & \{a,b,c\} \subseteq A. \end{cases}$$

We use $Q_{\mathrm{I}}, Q_{\mathrm{II}}, Q_{\mathrm{III}}$ and Q_{IV} to denote the set of quartet trees on leaf set $\{x,a,b,c\}$ in the case I, II, III and IV respectively, and let Q be the entire set of quartet trees for the leaf set of $\{x,a,b,c\}$. Since the four cases are mutually exclusive, Q_is partition $Q, i \in \{\mathrm{I},\mathrm{II},\mathrm{III},\mathrm{IV}\}$, and the sizes of Q_is are $|Q_{\mathrm{I}}| = \binom{k}{2} \times \binom{n-k}{1}$, $|Q_{\mathrm{II}}| = \binom{n-k}{3}$, $|Q_{\mathrm{III}}| = \binom{n-k}{2} \times \binom{k}{1}$ and $|Q_{\mathrm{IV}}| = \binom{k}{3}$.

Let $w(e)$ be a random variable of the weight that is added to e for a quartet tree of $\{x,a,b,c\}$. Consider $w(e)$ for each case $\{\mathrm{I},\mathrm{II},\mathrm{III},\mathrm{IV}\}$. Then we have:

- case I and III: $w(e) = \begin{cases} 1, & w.p. \frac{2}{3}; \\ 0 & w.p. \frac{1}{3}, \end{cases}$
- case II and IV: $w(e) = 0$.

Let $W_i(e), i \in \{\mathrm{I},\mathrm{II},\mathrm{III},\mathrm{IV}\}$, be the sum of all the weights added to the edge e. $W_{\mathrm{I}}(e)$ is a binomial random variable with parameters $\binom{k}{2}\binom{n-k}{1}$ and $\frac{2}{3}$; $W_{\mathrm{III}}(e)$ is a binomial random variable with parameters $\binom{n-k}{2}\binom{k}{1}$ and $\frac{2}{3}$; $W_{\mathrm{II}} = W_{\mathrm{IV}} = 0$. Let $W_n(e)$ be the sum of $W_i(e)$ values, so we have $W_n(e) = W_{\mathrm{I}}(e) + W_{\mathrm{III}}(e)$. Let $n_1 = \binom{k}{2}\binom{n-k}{1}$, and $n_2 = \binom{n-k}{2}\binom{k}{1}$, then

$$n_1 + n_2 = \frac{k(n-k)(n-2)}{2}$$

and so $W_n(e)$ consists of this many independent trials with probability of success on each trial of $\frac{2}{3}$. That is, $W_n(e)$ is a binomial random variable with parameters $\frac{k(n-k)(n-2)}{2}$ and $\frac{2}{3}$. □

Justification of Inequality (1)

Let E_n^{P} denote the set of pendent edges of current X_n-tree T_n, and E_n^{I} be the set of interior edges.

Lemma 4. *For any $e'' \in E_n^P$ and any $e' \in E_n^I$, the expected pendant edge total weight $W_n(e'')$ and the expected interior edge total weight $W_n(e')$, satisfy the inequality:*

$$\mathbb{E}[W_n(e')] - \mathbb{E}[W_n(e'')] \geqslant \frac{1}{3}\left[n^2 - 5n + 6\right] > 0. \tag{19}$$

Proof. $W_n(e'')$ and $W_n(e')$ are binomial random variables with the same probability of success $\frac{2}{3}$, but different number of trials $\binom{n-1}{2}$ and $\frac{k(n-k)(n-2)}{2}$, where $k \in \{2,\ldots,n-2\}$. Thus

$$\mathbb{E}[W_n(e'')] = \frac{2}{3}\binom{n-1}{2}, \quad \mathbb{E}[W_n(e')] = \frac{2}{3}\frac{k(n-k)(n-2)}{2}.$$

For a fixed n, $\mathbb{E}[W_n(e')] - \mathbb{E}[W_n(e'')]$ is a function of k. Therefore, to find the minimum of the difference between these two expected values, we need to find the value(s) of k for which $\mathbb{E}[W_n(e')] - \mathbb{E}[W_n(e'')]$ is minimal.

Let $y = (n-2)(n-k)k - (n^2-3n+2)$, then $\frac{dy}{dk} = (n-2)(n-2k)$. When $k = \frac{n}{2}, \frac{dy}{dk} = 0, \frac{d^2y}{dk^2} < 0$. Thus, there is a maximum at $k = \frac{n}{2}$, and minimum values occur at $k = 2$ or $k = n-2$. Therefore, when $k = 2$ or $k = n-2$,

$$\frac{1}{3}\left[n^2 - 5n + 6\right] \leqslant \mathbb{E}[W_n(e')] - \mathbb{E}[W_n(e'')]$$

Moreover, it is easily shown that for $n > 3, \frac{1}{3}\left[n^2 - 5n + 6\right] > 0$. Therefore,

$$\mathbb{E}[W_n(e')] - \mathbb{E}[W_n(e'')] \geqslant \frac{1}{3}\left[n^2 - 5n + 6\right] > 0. \qquad \square$$

Theorem 5. *For any $e'' \in E_n^P$ and any $e' \in E_n^I$,*

$$\mathbb{P}(W_n(e'') \geqslant W_n(e')) \leqslant 2\exp\left(-\frac{1}{144}n\right).$$

Proof. Let $W_n'' = W_n(e'') - \mathbb{E}[W_n(e'')], W_n' = W_n(e') - \mathbb{E}[W_n(e')]$, and $\beta = \mathbb{E}[W_n(e')] - \mathbb{E}[W_n(e'')]$. By Lemma 4, for $n \geqslant 4, \beta \geqslant 2dn^2$, where $d = \frac{1}{48}$. Now,

$$\begin{aligned} \mathbb{P}(W_n(e'') \geqslant W_n(e')) &= \mathbb{P}(W_n'' - W_n' \geqslant \beta), \\ &\leqslant \mathbb{P}\left(W_n'' \geqslant \frac{\beta}{2} \text{ or } -W_n' \geqslant \frac{\beta}{2}\right), \\ &\leqslant \mathbb{P}\left(W_n'' \geqslant \frac{\beta}{2}\right) + \mathbb{P}\left(-W_n' \geqslant \frac{\beta}{2}\right), \\ &\leqslant \mathbb{P}\left(W_n'' \geqslant dn^2\right) + \mathbb{P}\left(-W_n' \geqslant dn^2\right). \end{aligned}$$

We now apply Hoeffding's Inequality to the two terms on the right. Suppose that $\{Y_i, i = 1,2,3,\ldots,N\}$ are independent Bernoulli random variables, and let $Y = \sum_{i=1}^N Y_i$. By Hoeffding's Inequality [7], we have:

$$\mathbb{P}(Y - \mathbb{E}(Y) \geqslant t) \leqslant \exp\left(-2t^2/N\right),$$

$$\mathbb{P}(-(Y - \mathbb{E}(Y)) \geqslant t) \leqslant \exp\left(-2t^2/N\right).$$

Please cite this article in press as: S. Zhu, M. Steel, Does random tree puzzle produce Yule–Harding trees in the many-taxon limit?, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.02.003

Taking $Y = W'_n$ (and W''_n), $t = dn^2$, and $N = \frac{k(n-k)(n-2)}{2}$ in the previous string of inequalities, gives:

$$\mathbb{P}(W_n(e'') \geqslant W_n(e')) \leqslant 2\exp\left(-\frac{1}{576\frac{k}{n}(1-\frac{k}{n})(1-\frac{2}{n})}n\right) \leqslant 2\exp\left(-\frac{1}{144}n\right),$$

where the last inequality follows from $\frac{k}{n}(1-\frac{k}{n})(1-\frac{2}{n}) \leqslant \frac{1}{4}$, for $1 \leqslant k \leqslant n-1$. □

Justification of Inequality (2)

Proof. We will use Theorem 5 to establish Inequality (2). For $e'' \in E_n^{\mathrm{P}}$, and $e' \in E_n^{\mathrm{I}}$, let D be the event that $\min_{e'' \in E_n^{\mathrm{P}}}\{W_n(e'')\} < \min_{e' \in E_n^{\mathrm{I}}}\{W_n(e')\}$. Consider the complement of the event D,

$$D^c = \left(\min_{e \in E_n^{\mathrm{P}}}\{W_n(e'')\} < \min_{e' \in E_n^{\mathrm{I}}}\{W_n(e')\}\right)^c,$$

that is there is an interior edge e', such that $W_n(e') < \min_{e'' \in E_n^{\mathrm{P}}}\{W_n(e'')\}, W_n(e') \leqslant W_n(e''), \forall e'' \in E_n^{\mathrm{P}}$. Let $A_{e'',e'}$ be the event that $W_n(e'') > W_n(e')$, then we have, $D^c \subseteq \bigcup_{(e'',e') \in P \times I} A_{e'',e'}$, and so

$$\mathbb{P}(D^c) \leqslant \mathbb{P}\left(\bigcup_{(e'',e') \in P \times I} A_{e'',e'}\right).$$

According to Boole's inequality,

$$\mathbb{P}\left(\bigcup_{(e'',e') \in P \times I} A_{e'',e'}\right) \leqslant \sum_{(e'',e') \in P \times I} \mathbb{P}(A_{e'',e'}). \tag{20}$$

Now, the number of pendent edge is n, i.e. $|P| = n$, and the number of interior edge is $n-3$, i.e. $|I| = n-3$. Thus, $|P \times I| = n(n-3)$, and so, by Theorem 5, $\mathbb{P}(A_{e'',e'}) = \mathbb{P}(W_n(e'') \geqslant W_n(e')) \leqslant 2\exp(-\frac{1}{144}n)$. Thus,

$$\sum_{(e'',e') \in P \times I} \mathbb{P}(A_{e'',e'}) \leqslant n(n-3)2\exp\left(-\frac{1}{144}n\right) \leqslant 2n^2\exp\left(-\frac{1}{144}n\right). \tag{21}$$

Therefore,

$$\mathbb{P}\left(\min_{e'' \in E_n^{\mathrm{P}}}\{W_n(e'')\} \leqslant \min_{e' \in E_n^{\mathrm{I}}}\{W_n(e')\}\right) \geqslant 1 - 2n^2\exp\left(-\frac{1}{144}n\right). \quad \square$$

Justification of Inequality (3)

Proof. Since $\frac{k^2\exp(-ck)}{\exp(-ck/2)} = k^2\exp(-ck/2)$, and whenever $c \geqslant \frac{4\log k}{k}$ and $k > 1$, we have $k^2\exp(-ck/2) \leqslant 1$ we have:

$$k^2\exp(-ck) \leqslant \exp\left(-\frac{c}{2}k\right).$$

Thus, for $c \geqslant \frac{4\log k}{k}$ and $k > 1$, we have: $\sum_{k=m}^{\infty} k^2\exp(-ck) \leqslant \sum_{k=m}^{\infty}\exp\left(-\frac{c}{2}k\right)$.

Recognizing $\sum_{k=m}^{\infty}\exp\left(-\frac{c}{2}k\right)$ as the sum of a geometric series we have:

$$\sum_{k=m}^{\infty}\exp\left(-\frac{c}{2}k\right) = \frac{\exp(-cm/2)}{1-\exp(-c/2)}.$$

Thus, for $m \geqslant m_0, \exp(-cm/2) \leqslant \exp(-cm_0/2)$ and so, for $c \geqslant \frac{4\log k}{k}$ and $k > 1$, we have: $\sum_{k=m}^{\infty} k^2\exp(-ck) \leqslant \frac{\exp(-cm_0/2)}{1-\exp(-c/2)}$, as required. □

References

[1] A. Bagchi, A.K. Pal, Asymptotic normality in the generalized Polya–Eggenberger urn model, with an application to computer data structures, SIAM Journal on Algebraic and Discrete Methods 6 (1985) 394.
[2] V. Daubin, H. Ochman, Quartet mapping and the extent of lateral transfer in bacterial genomes, Molecular Biology and Evolution 1 (2004) 86.
[3] A. Dress, A. von Haeseler, M. Krueger, Reconstructing phylogenetic trees using variants of the "four-point condition", Studien zur Klassifikation 17 (1986) 299.
[4] J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, Journal of Molecular Evolution 17 (1981) 368.
[5] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, Systematic Biology 59 (2010) 307.
[6] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Systematic Biology 52 (2003) 696.
[7] W. Hoeffding, Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association 58 (1963) 13.
[8] C. Jordan, Sur les assemblages des lignes, Journal für die Reine und Angewandte Mathematik 70 (1869) 185.
[9] A.N.C. Kang, D.A. Ault, Some properties of a centroid of a free tree, Information Processing Letters 4 (1975) 18.
[10] A.F. Karr, Probability, Springer-Verlag, New York, 1993.
[11] S. Kotz, N. Balakrishnan, N.L. Johnson, Continuous multivariate distributions, Models and Applications, second ed., vol. 1, Wiley, New York, 2000.
[12] H. Mahmoud, Pólya Urn Models, Chapman and Hall/CRC, Boca Raton, 2008.
[13] A. McKenzie, M. Steel, Distributions of cherries for two models of trees, Mathematical Biosciences 164 (2000) 81.
[14] S.L. Mitchell, Another characterization of the centroid of a tree, Discrete Mathematics 24 (1978) 277.
[15] K. Nieselt-Struwe, A. von Haeseler, Quartet-mapping, a generalization of the likelihood-mapping procedure, Molecular Biology and Evolution 7 (2001) 1204.
[16] H.A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler, TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing, Bioinformatics 18 (2002) 502.
[17] C. Semple, M. Steel, Phylogenetics, Oxford University Press, Oxford, UK, 2003.
[18] R.T. Smythe, Central limit theorems for urn models, Stochastic Processes and their Applications 65 (1996) 115.
[19] M.A. Steel, Distributions on bicoloured binary trees arising from the principle of parsimony, Discrete Applied Mathematics 43 (1993) 245.
[20] K. Strimmer, N. Goldman, A. von Haeseler, Bayesian probabilities and quartet puzzling, Molecular Biology and Evolution 2 (1997) 210.
[21] K. Strimmer, A. von Haeseler, Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies, Molecular Biology and Evolution 13 (7) (1996) 964.
[22] L.S. Vinh, A. Fuehrer, A. von Haeseler, Random tree-puzzle leads to the Yule–Harding distribution, Molecular Biology and Evolution 28 (2011) 873.

Please cite this article in press as: S. Zhu, M. Steel, Does random tree puzzle produce Yule–Harding trees in the many-taxon limit?, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.02.003