Mathematical Biosciences 228 (2010) 90–99

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Priors for the Bayesian star paradox

Mikael Falconnet [*]

Université Joseph Fourier Grenoble 1, Institut Fourier UMR 5582 UJF-CNRS, 100 rue des Maths, BP 74, 38402 Saint Martin d'Hères, France

ARTICLE INFO

Article history:
Received 23 February 2010
Received in revised form 24 August 2010
Accepted 27 August 2010
Available online 8 September 2010

Keywords:
Phylogenetic trees
Bayesian statistics
Star trees

ABSTRACT

We show that the Bayesian star paradox, first proved mathematically by Steel and Matsen for a specific class of prior distributions, occurs in a wider context including less regular, possibly discontinuous, prior distributions.

1. Introduction

In phylogenetics, a particular resolved tree can be highly supported even when the data is generated by an unresolved star tree. This unfortunate aspect of the Bayesian approach to phylogeny reconstruction is called the *star paradox*. Recent studies highlight that the paradox can occur in the simplest nontrivial setting, namely for an unresolved rooted tree on three taxa and two states, see Yang and Rannala [7] and Lewis et al. [1]. Kolaczkowski and Thornton [2] presented some simulations and suggested that artifactual high posteriors for a particular resolved tree might disappear for very long sequences. Previous simulations in [7] were plagued by numerical problems, which left unknown the nature of the limiting distribution on posterior probabilities. For an introduction to the Bayesian approach to phylogeny reconstruction we refer to chapter 5 of Yang [5].

The statistical question which supports the star paradox is whether the Bayesian posterior distribution of the resolutions of a star tree becomes uniform when the length of the sequence tends to infinity, that is, in the case of three taxa and two states, whether the posterior distribution of each resolution converges to 1/3. In a recent paper, Steel and Matsen [3] disprove this, thus ruining Kolaczkowski and Thornton's hope, for a specific class of branch length priors which they call *tame*. More precisely, Steel and Matsen show that, for every tame prior and every fixed $\varepsilon > 0$, the posterior probability of any of the three possible trees stays above $1 - \varepsilon$ with non vanishing probability when the length of the sequence goes to infinity. This result was recognized by Yang [6] and reinforced by theoretical results on the posterior probabilities by Susko [4].

Our main result is that Steel and Matsen's conclusion holds for a wider class of priors, possibly highly irregular, which we call *tempered*. Recall that Steel and Matsen consider smooth priors whose densities satisfy some regularity conditions.

The paper is organized as follows. In Section 2, we describe the Bayesian framework of the star paradox. In Section 3, we define the class of tempered priors on the branch lengths and we state our main result. In Section 4, we state an extension of a technical lemma due to Steel and Matsen, which allows us to extend their result. In Section 5, we prove our main result. Section 6 is devoted to the proofs of intermediate results. In Appendix A, we prove that every tame prior, in Steel and Matsen's sense, is tempered, in the sense of this paper, and we provide examples of tempered, but not tame, prior distributions. Finally, in Appendix B, we prove the extension of Steel and Matsen's technical lemma stated in Section 4.

2. Bayesian framework for rooted trees on three taxa

We consider three taxa, encoded by the set $\tau = \{1,2,3\}$, with two possible states. Phylogenies on τ are supported by one of the four following trees: the star tree R_0 on three taxa and, for every taxon i in τ, the tree R_i such that i is the outlier. Relying on a commonly used notation, this reads as

$$R_1 = (1,(2,3)), \quad R_2 = (2,(1,3)), \quad R_3 = (3,(1,2)).$$

The phylogeny based on R_0 is specified by the common length of its three branches, denoted by t. For each i in τ, the phylogeny based on R_i is specified by a pair of branch lengths (t_e,t_ι), where t_e denotes the external branch length and t_ι the internal branch length, see Fig. 1.

[*] Tel.: +33 0 4 76 63 59 84; fax: +33 0 4 76 51 44 78.
E-mail address: mikael.falconnet@ujf-grenoble.fr

0025-5564/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2010.08.011

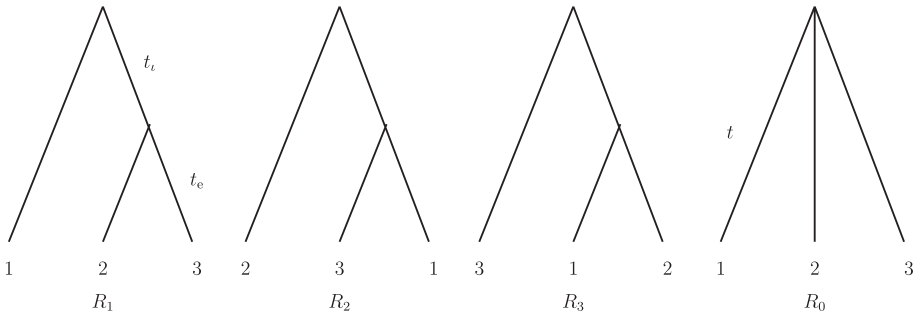

Fig. 1. The four rooted trees for three species.

For instance, in the phylogeny based on R_1, the divergence of taxa 2 and 3 occurred t_e units of time ago and the divergence of taxon 1 and a common ancestor of taxa 2 and 3 occurred $t_\iota + t_e$ units of time ago.

We assume that the sequences evolve according to a two-state continuous-time Markov process with equal substitution rates (which we may take to equal 1) between the two character states.

Four site patterns can occur. The first one, denoted by s_0, is such that a given site coincides in the three taxa. The three others, denoted by s_i with i in τ, are such that a given site coincide in two taxa and is different in the third taxon, which is taxon i. In other words, if one writes the site patterns in taxa 1, 2 and 3 in this order and x and y for any two different characters,

$$s_0 = xxx, \quad s_1 = yxx, \quad s_2 = xyx, \quad \text{and} \quad s_3 = xxy.$$

Let $\{s_0, s_1, s_2, s_3\}$ denote the set of site patterns in the specific case described above of three taxa and two states evolving in a two-state symmetric model. Assume that the counting of site pattern s_i is n_i. Then $n = n_0 + n_1 + n_2 + n_3$ is the total length of the sequences and, in the independent two-state symmetric model considered in this paper, the quadruple (n_0, n_1, n_2, n_3) is a sufficient statistics of the sequence data. We use the letter $\mathfrak{n}$ to denote any quadruple (n_0, n_1, n_2, n_3) of nonnegative integers such that $|\mathfrak{n}| = n_0 + n_1 + n_2 + n_3 = n \geqslant 1$.

For every site pattern s_i and every branch lengths (t_e, t_ι), let $p_i(t_e, t_\iota)$ denote the probability that s_i occurs on tree R_1 with branch lengths (t_e, t_ι). Standard computations provided by Yang and Rannala [7] show that

$$4p_0(t_e, t_\iota) = 1 + e^{-4t_e} + 2e^{-4(t_\iota + t_e)},$$
$$4p_1(t_e, t_\iota) = 1 + e^{-4t_e} - 2e^{-4(t_\iota + t_e)},$$
$$4p_2(t_e, t_\iota) = 4p_3(t_e, t_\iota) = 1 - e^{-4t_e}.$$

Let $\mathfrak{T} = (T_e, T_\iota)$ denote a pair of positive random variables representing the branch lengths (t_e, t_ι), and $\mathfrak{N} = (N_0, N_1, N_2, N_3)$ denote a quadruple of integer random variables representing the counts of sites patterns $\mathfrak{n} = (n_0, n_1, n_2, n_3)$.

3. The star tree paradox

Assuming that every taxon evolved from a common ancestor, the aim of phylogeny reconstruction is to compute the most likely tree R_i. To do so, in the Bayesian approach, one places prior distributions on the trees R_i and on their branch lengths $\mathfrak{T} = (T_e, T_\iota)$.

3.1. Main result

Let $\mathbb{P}(\mathfrak{N} = \mathfrak{n} | R_i, \mathfrak{T})$ denote the probability that $\mathfrak{N} = \mathfrak{n}$ assuming that the data is generated along the tree R_i conditionally on the branch lengths $\mathfrak{T} = (T_e, T_\iota)$. One may consider R_1 only since, for every $\mathfrak{n} = (n_0, n_1, n_2, n_3)$, the symmetries of the setting yield the relations

$$\mathbb{P}(\mathfrak{N} = \mathfrak{n} | R_2, \mathfrak{T}) = \mathbb{P}(\mathfrak{N} = (n_0, n_2, n_3, n_1) | R_1, \mathfrak{T}),$$

and

$$\mathbb{P}(\mathfrak{N} = \mathfrak{n} | R_3, \mathfrak{T}) = \mathbb{P}(\mathfrak{N} = (n_0, n_3, n_1, n_2) | R_1, \mathfrak{T}).$$

Notation 3.1. For every site pattern s_i, let P_i denote the random variable

$$P_i = p_i(\mathfrak{T}) = p_i(T_e, T_\iota).$$

For every i in τ and every $\mathfrak{n}$, let $\Pi_i(\mathfrak{n})$ denote the random variable

$$\Pi_i(\mathfrak{n}) = P_0^{n_0} P_1^{n_i} P_2^{n_j + n_k}, \quad \text{with} \quad \{i, j, k\} = \tau.$$

We recall that $P_2 = P_3$ and we note that, if $|\mathfrak{n}| = n_0 + n_1 + n_2 + n_3 = n$ with $n \geqslant 1$, then, for every i in τ,

$$\Pi_i(\mathfrak{n}) = P_0^{n_0} P_1^{n_i} P_2^{n - n_0 - n_i}.$$

Fix $\mathfrak{n}$ and assume that $|\mathfrak{n}| = n_0 + n_1 + n_2 + n_3 = n$ with $n \geqslant 1$. For every i in τ, the posterior probability of R_i conditionally on $\mathfrak{N} = \mathfrak{n}$ is

$$\mathbb{P}(R_i | \mathfrak{N} = \mathfrak{n}) = \frac{n!}{n_0! n_1! n_2! n_3!} \frac{1}{\mathbb{P}(\mathfrak{N} = \mathfrak{n})} \mathbb{E}(\Pi_i(\mathfrak{n})).$$

Thus, for every i and j in τ,

$$\frac{\mathbb{P}(R_i | \mathfrak{N} = \mathfrak{n})}{\mathbb{P}(R_j | \mathfrak{N} = \mathfrak{n})} = \frac{\mathbb{E}(\Pi_i(\mathfrak{n}))}{\mathbb{E}(\Pi_j(\mathfrak{n}))}.$$

For every $\varepsilon > 0$ and every i in τ, let $\mathcal{N}_i^\varepsilon$ denote the set of $\mathfrak{n}$ such that, for both indices j in τ such that $j \neq i$,

$$\mathbb{E}(\Pi_i(\mathfrak{n})) \geqslant (2/\varepsilon) \mathbb{E}(\Pi_j(\mathfrak{n})).$$

One sees that, for every i in τ and $\mathfrak{n}$ in $\mathcal{N}_i^\varepsilon$,

$$\mathbb{P}(R_i | \mathfrak{N} = \mathfrak{n}) \geqslant 1 - \varepsilon,$$

which means that the posterior probability of tree R_i among the three possible trees is highly supported.

Recall that, under hypothesis R_0 and for a tame prior distribution on $\mathfrak{T} = (T_e, T_\iota)$, Steel and Matsen prove that, for every i in τ, $\mathbb{P}(\mathfrak{N} \in \mathcal{N}_i^\varepsilon)$ does not go to *0* when the sequence length n goes to infinity, and consequently that the posterior probability $\mathbb{P}(R_i | \mathfrak{N})$ can be close to 1 even when the sequence length n is large.

As stated in the introduction, our aim is to prove the same result for tempered prior distributions of $\mathfrak{T} = (T_e, T_\iota)$, which we now define.

Notation 3.2.

(1) For every $s \in [0, 1]$ and $z \in [0, 3]$, let

$$G(z, s) = \mathbb{P}\left(e^{-4T_e}(1 - e^{-4T_\iota}) \leqslant s \,|\, e^{-4T_e}(1 + 2e^{-4T_\iota}) = z\right).$$