The star-tree paradox in Bayesian phylogenetics

Bengt Autzen

Department of Philosophy, Logic and Scientific Method

LSE

Overview

1) Introduction
2) The facts
3) The (alleged) paradox
4) The star-tree paradox and the meaning of posterior probabilities of trees
5) Symmetry

1. Introduction

What is the ‘star-tree paradox’ about?

“The star-tree paradox refers to the conjecture that the posterior probabilities for […] the three rooted trees for three species […] do not approach 1/3 when the data are generated using the star tree and when the amount of data approaches infinity.” (Yang, 2007)

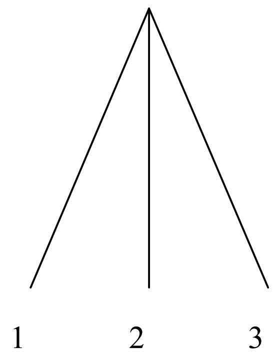
2. The facts

2.1 Phylogenetic estimation problem given three species

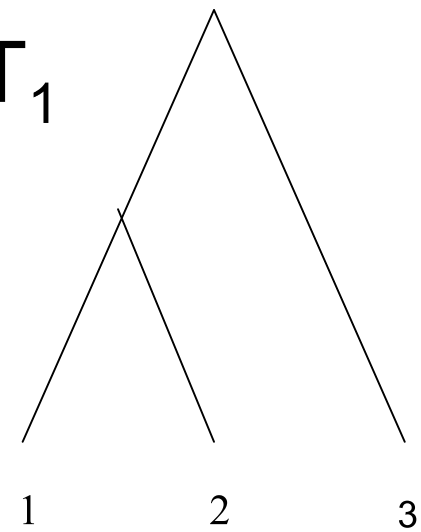
a) The tree topologies:

Star-tree T_0 and three binary trees T_1, T_2, and T_3

T_0



T_1



2. The facts (cont.)

b) The (synthetic) data:

Three DNA sequences, n nucleotides long, nucleotides are binary characters.

Hence, 2^3 = 8 possible data configurations at a nucleotide site (‘site pattern’) or four site patterns xxx, xxy, yxx, and xyx, where x and y are any two different nucleotides.

Data are summarized as counts of these four site patterns n_0, n_1, n_2, and n_3.

2. The facts (cont.)

c) Model of nucleotide substitution:

2-state symmetric Markov process

Probabilities of site patterns under tree T_1:

$$p_0(t_0,t_1) = \frac{1}{4} + \frac{1}{4}e^{-4t_1} + \frac{1}{2}e^{-4(t_0+t_1)},$$

$$p_1(t_0,t_1) = \frac{1}{4} + \frac{1}{4}e^{-4t_1} - \frac{1}{2}e^{-4(t_0+t_1)},$$

$$p_2(t_0,t_1) = \frac{1}{4} - \frac{1}{4}e^{-4t_1} = p_3(t_0,t_1).$$

2. The facts (cont.)

d) Likelihood function for tree T_1 (with proportionality constant C):

$$P(n_0, n_1, n_2, n_3 \mid T_1, t_0, t_1)$$

$$= Cp_0^{n_0} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

$$= Cp_0^{n_0} p_1^{n_2} p_2^{n_2 + n_3}.$$

2. The facts (cont.)

Similarly for trees T_2 and T_3:

$$P(n_0, n_1, n_2, n_3 \mid T_2, t_0, t_1)$$
$$= C p_0^{n_0} p_1^{n_2} p_2^{n_3 + n_1}$$
$$P(n_0, n_1, n_2, n_3 \mid T_3, t_0, t_1)$$
$$= C p_0^{n_0} p_1^{n_3} p_2^{n_1 + n_2}$$

2. The facts (cont.)

e) Prior probabilities:

The three binary trees T_1, T_2 and T_3 have equal prior probability 1/3. Hence, the star tree T_0 gets assigned 0 prior probability.

The prior distribution on branch lengths t_0, t_1 is the same for each tree with a smooth joint probability density function that is bounded and everywhere nonzero (e.g. exponential prior (Yang and Rannala (2005)).

2. The facts (cont.)

2.2 Steel and Matsen's theorem (Steel and Matsen (2007)):

Consider sequences of length n generated by the star-tree with strictly positive edge length t and let n_0, n_1, n_2, and n_3 be the resulting data (in terms of site patterns).

Further, the aforementioned assumptions regarding the process of nucleotide substitution and the prior probability distributions hold. Then…

2. The facts (cont.)

Steel and Matsen's theorem (cont.):

For any $\varepsilon>0$, and each binary tree T_i ($i=1,2,3$), the probability that n_0, n_1, n_2, and n_3 has the property that

$$P(T_i|n_0, n_1, n_2, n_3) > 1- \varepsilon$$

does not converge to 0 as n tends to infinity.

2. The facts (cont.)

2.3 Simulation Results Yang (2007):

For data sets of size n = 3*10^9 simulated under the star-tree the posterior probability distribution of the three binary trees fails to form a uniform distribution (1/3, 1/3, 1/3) for several data sets. That is, at least one of the three posterior probabilities is > 0.95 in 4.23% of data sets, and in 0.79% of data sets at least one of the three posterior probabilities is > 0.99. In 17.3% of data sets at least one of the three posterior probabilities is <0.05 and in 2.6% of data sets at least one of the three posterior probabilities is <0.01.

3. The (alleged) paradox

Question: What is paradoxical about the ‘star-tree paradox’?

Steel and Matsen’s theorem as well as simulation results are in conflict with Yang’s criteria which a ‘reasonable’ Bayesian method should satisfy…

3. The (alleged) paradox

Yang's criteria (Yang, 2007):

1) The posterior probabilities of the three binary trees converge to the uniform distribution (1/3, 1/3, 1/3) when n tends to infinity if the 'true' tree is the star tree and only the three binary trees get assigned positive priors.
2) If a binary tree is the true tree, its posterior probability should converge to 1 when n tends to infinity.

3. The (alleged) paradox

How to justify Yang’s criteria?

Maybe they follow from the meaning of posterior probabilities of trees?

4. The star-tree paradox and the meaning of posterior probabilities of trees

A suggested interpretation of PP of trees:

"We use the case where the full model is correct – that is, where the analysis model matches the simulation model – to illustrate the interpretation of posterior probabilities for trees. When the data are simulated under the prior and when the full analysis model is correct, the posterior for a tree is the probability that the tree is true." (Yang and Rannala, 2005, p. 457)

4. The star-tree … (cont.)

What do YR mean by ‘probability that a tree is correct’?

For a given tree with PP x, the frequency that the PP of the true tree (i.e. data generating tree) in an interval of length 0.2 containing PP x is called the ‘probability that the tree with PP x is correct’.

4. The star-tree … (cont.)

Example:

Trees with PP between 0.94 and 0.96 have all PP close to 0.95. Among them, about 95% are the posterior probabilities of the true tree while others (about 5%) are posterior probabilities for one of the two incorrect trees.

4. The star-tree … (cont.)

Problem with YR’s interpretation of PP:

In the case of criterion 1) the simulation and the analysis model do not match! That is, the star-tree topology gets zero prior in the analysis model.

The relation between PP of a tree and what Yang and Rannala call ‘probability that the tree is correct’ is an empirical phenomenon, not a conceptual necessity.

4. The star-tree … (cont.)

Where does this leave us regarding the meaning of PP of trees?

Prior and posterior probabilities as a (subjective) degrees of belief?

5. Symmetry

A further justification for Yang's criterion 1) might come from symmetry considerations.

Aren't the three binary trees – in an intuitive way - equally similar (or dissimilar) to the star tree?

5. Symmetry

However, why should the symmetry of the problem result in the convergence of the PP of trees to the uniform distribution (1/3, 1/3, 1/3)? There are symmetries to be found in behaviour of the PP for trees when n tends to infinity, but they are of a different kind (see Matsen/Steel's theorem).

References

Steel, M. and Matsen, F. (2007): ‘The Bayesian “Star Paradox” Persists for Long Finite Sequences’, in *Mol. Biol. Evol.* 24(4)

Yang, Z. (2007): ‘Fair-Balance Paradox, Star-tree Paradox, Bayesian Phylogenetics’, in *Mol. Biol. Evol.* 24(8)

Yang, Z. and Rannala, B. (2005): ‘Branch-Length Prior influences Bayesian posterior Probability of Phylogeny’, in *Syst. Biol.* 54(3)