A starless bias in the maximum likelihood phylogenetic methods

Xuhua Xia[1][2]

[1]University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

[2]Ottawa Institute of Systems Biology, Ottawa, Canada K1H 8M5

Abstract

The maximum likelihood method is the gold standard in molecular phylogenetics and accounts for nearly half of all published phylogenetic trees, but the method has a strange phylogenetic bias so far not explored. If the aligned sequences are equidistant from each other with the true tree being a star tree, then the likelihood method is incapable of recovering it unless the sequences are either identical or extremely diverged. Here I analytically demonstrate this "starless" bias and identify the source for the bias. In contrast, distance-based methods (with the least-squares method for branch evaluation and either minimum evolution or least-squares criterion for choosing the best tree) do not have this bias. The finding sheds light on the star-tree paradox in Bayesian phylogenetic inference.

Key Words: maximum likelihood, molecular phylogenetics, distance-based phylogenetic method, starless, star-tree paradox

1. Illustration of the bias

The starless bias refers to the bias of a phylogenetic method that cannot recover a star tree even when the star tree is the true tree. It was first alluded to in a study of potential bias in maximum likelihood method involving missing data and rate heterogeneity over sites (Xia 2014). I will illustrate this bias here numerically, identify the source of the bias, and discuss its relevance to the star-tree paradox associated with Bayesian phylogenetic inference (Lewis *et al.* 2005; Yang 2007; Yang and Zhu 2018).

Suppose we have a set of four sequences (Figure 1A) generated from the JC69 substitution model (Jukes and Cantor 1969). This simplest substitution model, together with other frequently used Markovian nucleotide substitution models such as F84 (used in DNAML since 1984, Hasegawa and Kishino 1989; Kishino and Hasegawa 1989), HKY85 (Hasegawa *et al.* 1985), TN93 (Tamura and Nei 1993), and GTR (Lanave *et al.* 1984; Tavaré 1986) have been numerically illustrated in great detail (Xia 2017; Xia 2018b). The four sequences have the same nucleotide frequencies. The last four sites are all identical, but the first four sites differ, with each sequence differing from the other three by exactly four nucleotide substitutions. There are twice as many transversions as transitions as one would expect from JC69. One may note that some sequences differ from others by two transitions and two transversions (e.g., S1 and S2), or by four transversions (e.g., between S1 and S4). However, JC69 does not discriminate between transitions and transversions. If we impose the JC69 model, then the four sequences are expected to be equidistant from each other (with D_{ij} = 0.82396), and should be related by a star tree (Figure 1B), with branch lengths being $D_{ij}/2$ from the internal node to each of the four leaves. Indeed, a distance method such as Neighbor-Joining (Saitou and Nei 1987) or FastME (Desper and Gascuel 2002; 2004) will recover the expected star tree in Figure 1B. In fact, we would expect all reasonable phylogenetic reconstruction method to recover the star tree in Figure 1B.

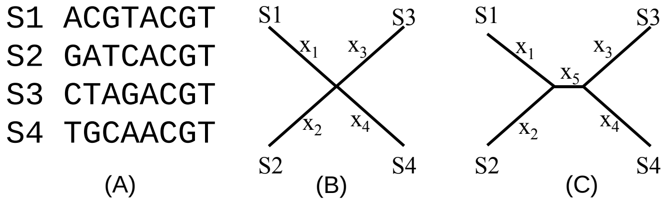

Figure 1: Phylogenetic bias in ML. (A) Four sequences conforming the JC69 model. (B) A star tree. (C) ML tree with $x_5 > 0$.

Surprisingly, the ML tree (Figure 1C), given the sequence data in Figure 1A and the JC69 model, has $x_5 > 0$, with its lnL = -39.843371588, achieved when $x_1 = x_2 = x_3 = x_4 = 0.482505887$ and $x_5 = 0.133851489$. Furthermore, the three alternative unrooted trees all have exactly the same branch lengths and the same lnL. Thus, none of three ML trees is correct. While the difference in lnL between the Figure 1B tree and the Figure 1C tree is small with eight sites in Figure 1A, the difference increases proportionally with sequence length.

One may argue that the ML method does not do anything wrong because all three possible topologies are equally supported. However, in practical phylogenetic analysis, the ML method does not generate all possible topologies but instead produces just one. One might also argue that sites 1-4 and sites 5-8 in Figure 1A represent extreme rate heterogeneity, so the sequence alignment is unrealistic. However, one may obtain the same results with a large number of sites with intermediate variability over sites. My choice of the eight sites are simply for easy illustration. The starless bias does not need to have such two groups of sites with extreme rate heterogeneity over sites.

One might suspect that the likelihood function may reach a local maximum, so the tree in Figure 1C is not the true ML tree. However, this is not the case. The lnL value will decrease if we force $x_5 = 0$ and re-optimize branch lengths. With the constraint of $x_5 = 0$, the best lnL is -39.878988659, achieved when $x_1 = x_2 = x_3 = x_4 = 0.524860294$. This lnL is smaller than that for the ML tree with $x_5 > 0$.

One may also argue that the problem is caused by misspecification of the substitution model. The aligned sequences in Figure 1A has four highly variable sites and four invariant sites, and therefore exhibit a high rate heterogeneity over sites which is not accommodated by the aforementioned likelihood calculation assuming a Poisson-distributed rate. However, this argument does not remove the problem because it is perfectly easy to add nucleotide sites with intermediate variation so that a likelihood ratio test or an information theoretic index (Burnham and Anderson 2002; Xia 2009) would prefer the Poisson-distributed rate model over the more complicated rate heterogeneity model such as gamma-distributed rates. We will find the problem remains with such new data sets that do not demand a model with rate heterogeneity over sites.

Thus, when the star tree is the true tree, likelihood-based method will take one branch from the root node and stick it somewhere on another branch, leading to three wrong but equally supported trees (with one shown in Figure 1C). In short, any of the three equally supported trees is wrong. This example also suggests that the ML method may favor certain tree shapes relative to distance-based methods. This topic does not seem to have been studied.

To facilitate exposition, we will designate the tree with $x_5 > 0$ as $T_{x5>0}$ and the tree with $x_5 = 0$ as $T_{x5=0}$, with their corresponding log-likelihood (lnL) as $\ln L_{x5>0}$ and $\ln L_{x5=0}$, respectively. Given the optimized branch lengths for $T_{x5>0}$ and $T_{x5=0}$, the tree length is shorter for $T_{x5>0}$ than that for $T_{x5=0}$.

The lnL value for the star tree obtained by the distance method, with $x_i = 0.41198$, has lnL = -40.13321515, which is worse based on the maximum likelihood criterion. Thus, the likelihood principle gives us a wrong tree with inexplicably weird branch lengths. I have checked the likelihood calculation by using both the pruning algorithm (Felsenstein 1973; 1981) and the brute-force approach by writing down all 16 terms for the tree in Figure 1C. The pruning algorithm for four sequences has been numerically illustrated before (Xia 2018a). Likelihood for the first and fifth sites (designated L_1 and L_5, respectively), given the tree in Figure 1C and the brute-force approach, are

$$\begin{aligned} L_1 = {} & \pi_A P_{AA}(x_1)P_{AG}(x_2)P_{AA}(x_5)P_{AC}(x_3)P_{AT}(x_4) \\ & +\pi_C P_{CA}(x_1)P_{CG}(x_2)P_{CA}(x_5)P_{AC}(x_3)P_{AT}(x_4) \\ & +\ldots+\pi_T P_{TA}(x_1)P_{TG}(x_2)P_{TT}(x_5)P_{TC}(x_3)P_{TT}(x_4) \end{aligned} \tag{1}$$

$$\begin{aligned} L_5 = {} & \pi_A P_{AA}(x_1)P_{AA}(x_2)P_{AA}(x_5)P_{AA}(x_3)P_{AA}(x_4) \\ & +\pi_C P_{CA}(x_1)P_{CA}(x_2)P_{CA}(x_5)P_{AA}(x_3)P_{AA}(x_4) \\ & +\ldots+\pi_T P_{TA}(x_1)P_{TA}(x_2)P_{TT}(x_5)P_{TA}(x_3)P_{TA}(x_4) \end{aligned} \tag{2}$$

where π_A, π_G, π_C, and π_T are equilibrium frequencies and all equal to 0.25, P_{ii} and P_{ij} are transition probabilities involving two identical or different nucleotides, respectively, in two neighboring nodes, x_i are branch lengths as indicated in Figure 1C. The derivation for the JC69, as well as for F84, HKY85, TN93, and GTR models, have been numerically illustrated in great detail (Xia 2017; Xia 2018b).

Given the JC69 model, there are only two site patterns in the aligned sequences in Figure 1A, one shared among the first four site with the same likelihood as L_1, and the other shared among the last four sites with the same likelihood as L_5. The JC69 model implies that the four P_{ii} functions are the same, so are the 12 P_{ij} functions. With $x_5 = 0$, the 16 terms in L_1 and L_5 are reduced to only four terms where the two internal nodes are occupied by the same nucleotide. This is because $P_{ii}(0) = 1$ and $P_{ij}(0) = 0$. One can readily replicate numerically this bias which was aptly termed "the starless bias" by Sudhir Kumar (pers. comm.).

2. The source of the bias

Why does the ML method arrive at a tree with $x_5 = 0.13385149$ instead of $x_5 = 0$? The underlying cause becomes clear if we take notice of two things. First, the tree length (the summation of all branches of a tree) is longer for the Figure 1C tree than for the Figure 1B tree (2.063875 vs 1.647920). Second, highly variable nucleotide sites 1-4 in Figure 1A would favor longer branches and their site-specific likelihood values (represented by L_1) will be greater for the Figure 1C tree (with longer tree length) than for the Figure 1B tree (with shorter tree length). In contrast, the invariant nucleotide sites 5-8 will favor short branches and their site-specific likelihood values (represented by L_5) will be greater for the Figure 1B tree (with shorter tree length) than for the Figure 1C tree (with longer tree lengths). Thus, L_1 is 0.001233825 for Figure 1C tree, but 0.000805733 for the Figure 1B tree. This is equivalent to say that variable nucleotide sites 1-4 in Figure 1A favor the Figure 1C tree with longer tree lengths than the Figure 1B tree with shorter branches. As a matter of fact, such extremely variable sites would favor a tree with infinitely long branches. In contrast, L_5 is 0.038265498 for the Figure 1C tree, but 0.054500457 for the Figure 1B tree. This means that the four invariant nucleotide sites 5-8 favor the shorter Figure 1B tree than the longer Figure 1C tree. Such sites indeed would favor zero-length branches.

The tree lnL values for the Figure 1C tree ($\ln L_{x5>0}$) and the Figure 1B trees ($\ln L_{x5=0}$) are, respectively,

$$\ln L_{x5>0} = 4\ln(L_1) + 4\ln(L_5) = -39.843371588 \tag{3}$$

$$\ln L_{x5=0} = 4\ln(L_1) + 4\ln(L_5) = -40.13321515 \tag{4}$$

There are only two special cases where a star tree will be reconstructed. The first is when all nucleotide sites are invariant like sites 5-8 in Figure 1A, and the two trees will converge to a star tree with zero-length branches. The second is when all sites are extremely variable like sites 1-4 when the two trees will converge to a star tree with infinitely long branches. In practical molecular phylogenetics, if we have far more variable sites than invariant sites, or far more invariant sites than variable sites, then the two trees will have their respective lnL approaching each other. For four sequences, if we fix the number of invariant sites to10 and increase the number of variable sites to greater than 35, then the two topologies will have the same lnL up to 6 digits after decimal point. Whenever we have a roughly equal mixture of the two categories of sites, the ML method will miss the star tree even if it is the true tree. Such a mixture of sites is equivalent to rate heterogeneity over sites. Thus, the finding here may help explain the phylogenetic distortion involving rate heterogeneity (Kuhner and Felsenstein 1994; Xia 2014).

3. Discussion

This starless bias associated with the ML method sheds light on the well-known star-tree paradox in which Bayesian phylogenetic inference prefers one of the three topologies when the true tree is a star tree (Lewis *et al.* 2005; Yang 2007; Yang and Zhu 2018). The finding reported here suggests that the problem may not be caused by Bayesian inference but instead is caused by the ML method that favors a resolved tree against the star tree. It also suggests that the two proposed solutions (Lewis *et al.* 2005; Yang 2007) are unlikely to resolve the problem. The first solution of assigning nonzero prior probability for the degenerate star tree (Lewis *et al.* 2005; Yang 2007) will not work because the nonzero prior probability assigned to the star tree will eventually be offset by the likelihood difference favoring the resolved tree when sequence length increases to infinity. The second solution of increasing informative prior forcing the internal branch length towards zero (Yang 2007) will have the same problem, unless the informative prior increases with sequence length. The reversible-jump Markov chain Monte Carlo algorithm proposed by Lewis et al. (2005), albeit ingenious, is unlikely to solve the problem because the starless bias is not due to the star tree being excluded from tree searching but because it has a smaller likelihood value than any of the three resolved tree. The star tree will be increasingly disfavored by increasing amount of data.

In short, if the true tree is a star tree, then there are only two special cases in which the likelihood method will recover the star tree. One is a trivial case when all sequences are identical (Xia 2014). The other is when sequences are all highly and equally diverged. The star tree cannot be recovered other than these two extreme cases. The results also suggest that the maximum likelihood and the distance-based methods may favor different tree shapes.

Acknowledgements

This study is supported by Discovery Grant from Natural Science and Engineering Research Council (NSERC, RGPIN/ 2018-03878) of Canada. I thank Guy Baele, Sudhir Kumar, Laura Kubatko and Arindam RoyChoudhury for discussion.

References

Burnham KP, Anderson DR. 2002. Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach. New York, NY: Springer.

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9(5):687-705.

Desper R, Gascuel O. 2004. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol. Biol. Evol. 21(3):587-98.

Felsenstein J. 1973. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240-249.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Hasegawa M, Kishino H. 1989. Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. Jpn J Genet 64(4):243-58.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22(2):160-74.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21-123.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. 29:170-179.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11(3):459-68.

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20(1):86-93.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. Syst Biol 54(2):241-53.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512-526.

Tavaré S. 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. In: Miura RM, editor. Some mathematical questions in biology – DNA sequence analysis. Providence, RI: American Mathematical Society. p. 57-86.

Xia X. 2009. Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. Mol. Phylogenet. Evol. 52:665-676.

Xia X. 2014. Phylogenetic Bias in the Likelihood Method Caused by Missing Data Coupled with Among-Site Rate Variation: An Analytical Approach. In: Basu M, Pan Y, Wang J, editors. Bioinformatics Research and Applications.: Springer. p. 12-23.

Xia X. 2017. Deriving Transition Probabilities and Evolutionary Distances from Substitution Rate Matrix by Probability Reasoning. J Genet Genome Res 3:031.

Xia X. 2018a. Maximum Likelihood in Molecular Phylogenetics. Bioinformatics and the Cell. Springer, Cham. p. 381-395.

Xia X. 2018b. Nucleotide Substitution Models and Evolutionary Distances. Bioinformatics and the Cell. Springer, Cham. p. 269-314.

Yang Z. 2007. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol Biol Evol 24(8):1639-55.

Yang Z, Zhu T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. Proc. Natl. Acad. Sci. U S A 115(8):1854-1859.