Phylogenetic Bias in the Likelihood Method Caused by Missing Data Coupled with Among-Site Rate Variation: An Analytical Approach

Xuhua Xia

Department of Biology and Center for Advanced Research in Environmental Genomics,
University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A,
Ottawa, Ontario, Canada, K1N 6N5
xxia@uottawa.ca

Abstract. More and more researchers in phylogenetics are concatenating gene sequences to produce supermatrices in the hope that larger data sets will lead to better phylogenetic resolution. Almost all of these supermatrices contain a high proportion of missing data which could potentially cause phylogenetic bias. Previous studies aiming to identify the missing-data-mediated bias in the maximum likelihood method have noted a bias associated with among-site rate variation. However, this finding is by sequence simulation and has been challenged by other simulation studies, with the controversy still unresolved. Here I illustrate analytically this bias caused by missing data coupled with among-site rate variation. This approach allows one to see how much the bias can contribute to likelihood differences among different topologies. The study highlights the point that, while supermatrices may lead to "robust" trees, such "robust" trees may be purchased with illegal phylogenetic currency.

Keywords: missing data, pruning algorithm, likelihood, phylogenetic bias, supermatrix.

1 Introduction

Many supermatrices have been compiled in recent years by concatenating sequences from many different genes [1-4]. Such concatenated genes typically have few shared sites among all included species. For example, while Regier et al. [3] claimed to have 41 kilobases of aligned DNA sequences, the actual number of sites that are completely unambiguous among all 80 species amounts to only 705 sites. Some genes are completely missing in nearly half of the 80 species. While the potential problems involving such "?"-laden supermatrices have been suspected before[5], specific biases associated with such missing data have not been well studied, especially not in the likelihood framework which has been the gold standard in phylogenetic reconstruction.

Previous studies [6-11] attempted to identify bias associated with missing data either by sequence simulation or by selectively eliminating sites in a real sequence alignment. While most publications suggest that phylogenetic reconstruction is not sensitive to missing data or that the benefit of including taxa with missing data

M. Basu, Y. Pan, and J. Wang (Eds.): ISBRA 2014, LNBI 8492, pp. 12–23, 2014.

out-weight the cost of their exclusion [6, 8-11], a recent study [7] suggested a significant bias associated with missing data and coupled with among-site rate variation. However, such simulation-based findings often cannot pin-point where the bias arises and consequently have been challenged by others on both empirical [6, 9, 11] and theoretical grounds [9], although these latter publications did not explicitly test the claimed bias [7] associated with among-site variation. Roure et al. [9] noted that, if sequences contain similar phylogenetic information, then phylogenetic reconstruction is not sensitive to missing data. However, they also noted that heterogeneous data could lead to phylogenetic bias based on extensive data analysis.

Here I demonstrate analytically the bias associated with the missing data coupled with among-site rate variation. The pruning algorithm [12, 13, 14, pp. 253-255] is briefly outlined, in conjunction with the conventional missing data handling by the likelihood method, so that the reader can verify the claimed bias introduced by missing data. I first illustrate the "bias" shown by Lemmon et al. [7] when branch lengths are not allowed to be zero, by using both JC69 [15] and F84 [16] models. Such a "bias" can be easily avoided by simply allow branches to be zero and should not be considered as estimation bias in the likelihood method. However, the bias due to the missing data associated with among-site rate variation [7] is real. This bias can lead to either increased tendency (and confidence) to group together OTUs (operational taxonomic units) that share the same stretches of missing sites or in the opposite direction. The results suggest that blindly concatenating sequence data to generate a supermatrix with many pieces of missing data will generate false confidence in phylogenetic resolution and should be avoided.

2 Missing Data Handling and the Pruning Algorithm

The likelihood approach features a convenient way to handle missing data, which is best illustrated with the pruning algorithm. Suppose we have four OTUs with sequence data in Fig. 1a, and with the last two sequences being entirely missing (represented by '?'). Obviously, we can only estimate the distance between S1 and S2 but not the evolutionary relationships involving OTUs S3 or S4. The maximum likelihood distance between S1 and S2, based on the JC69 model, is given by

$$L = \frac{8!}{4!4!} P_{ii}^4 P_{ij}^4 \tag{1}$$

which, when maximized, leads to a distance of 0.8239592165.

Fig. 2 illustrates the computation of the likelihood by the pruning algorithm, given the first site of the aligned nucleotide sequence (Fig. 1a) and topology T_1 in Fig. 1. I included the numerical illustration here to facilitate the verification of subsequent claims that the maximum likelihood method does exhibit a true and identifiable bias in phylogenetic reconstruction involving missing data coupled with among-site rate variation.

Fig. 1. Three sets of sequences, (a), (b) and (c), for four OTUs (operational taxonomic units), and three alternative topologies (T1, T2 and T3) for illustrating phylogenetic bias introduced by missing data. Branch lengths are represented by bi. The sequences in bold italic in (c) are the same as those in (b). Note that the three variable sites at the 5' end could be diffused at different sites in the data instead of clumping together to be so easily recognized.

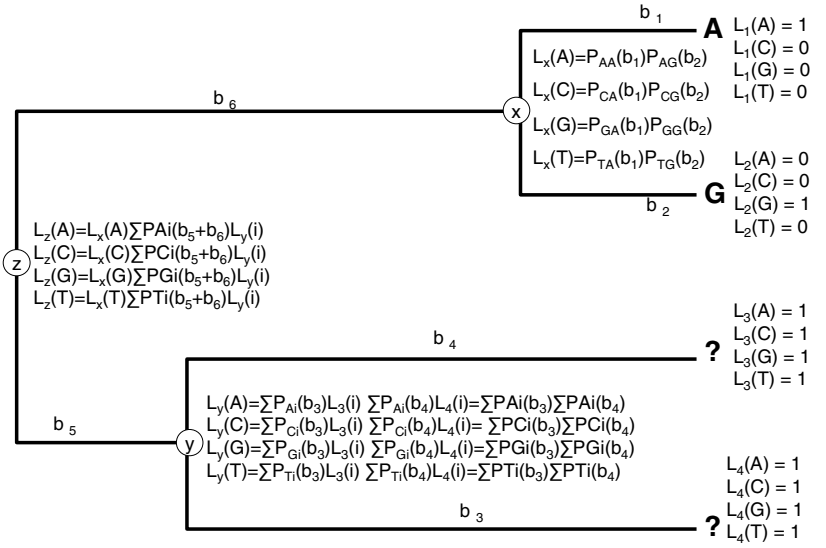

Fig. 2. Likelihood computation with the pruning algorithm [14, pp. 253-255]

We first define an array for each of the nodes including the leaf nodes. The array contains four elements for nucleotide sequences and 20 for amino acid sequences. For a leaf node i with a resolved nucleotide S, $L_i(S) = 1$, and $L_i(\text{not S}) = 0$. For an unknown or missing nucleotide, $L_i(1) = L_i(2) = L_i(3) = L_i(4) = 1$. For an internal node i with two offspring (o_1 and o_2), L_i is recursively defined as

$$L_i(s) = \left[\sum_{k=0}^{3} P_{sk}(b_{i,o_1}) L_{o_1}(k)\right]\left[\sum_{k=0}^{3} P_{sk}(b_{i,o_2}) L_{o_2}(k)\right] \tag{2}$$

where $b_{i,o1}$ means the branch length between internal node i and its offspring o_1, and P_{sk} is the transition probability from state s to state k. For example, $b_{x,S1}$ (branch length between internal node x and its offspring S_1) is b_1 in Fig. 2. Internal node z is special in that we cannot estimate b_5 and b_6 separately because the resulting tree is unrooted. We simply move node z to the location of node y (or node x), so that either b_5 or b_6 is 0 and the other is then equal to (b_5+b_6). If b_5 is 0, then $P_{ii}(b_5) = 1$ and $P_{ij}(b_5) = 0$, i.e., no time for anything to change. This leads to the simplified equations for computing $L_z(i)$ in Fig. 2. The final likelihood is

$$L = \sum_{i=1}^{4} \pi_i L_z(i) \tag{3}$$

where π_i is the frequency of nucleotide i.

Given the JC69 model, the sequences in Fig. 1a have two site patterns, with the first four sites sharing one site pattern and the last four sites sharing the other site pattern. Designating the likelihood of the two site patterns in Fig. 1a as $L_{a.pattern1}$ and $L_{a.pattern2}$, the log-likelihood (lnL) for all eight sites (Fig. 1a), given topology T_1 in Fig. 1), is

$$\ln L = 4\ln(L_{a.pattern1}) + 4\ln(L_{a.pattern2}) \tag{4}$$

which, upon maximization, leads to $b_1 + b_2 = 0.8239592165$, and lnL = -21.02998149. This is perfectly consistent with the result from Eq. (1) as we would have expected. Terms including b_3, b_4 and b_5+b_6 all cancel out in Eq. (4), suggesting that the sequences in Fig. 1a have absolutely no information for estimating b_3, b_4 and b_5+b_6, which again is what we would have expected. Note that lnL would be greater if we treat the two site patterns as two separate partitions and estimate branch lengths separately. Assuming the JC69 model, the maximum likelihood is 0.25^2 for each site in the first partition (reached when b_1 and b_2 are infinitely large) and 0.25 for each site in the second partition (reached when $b_1 = b_2 = 0$), so lnL will then be

$$\ln L = 4\ln(0.25^2) + 4\ln(0.25) = -16.63553233 \tag{5}$$

which indicates that maximizing lnL by partitioning the data may not be a good idea given the dramatically incompatible branch length estimates from the two partitions.

If we perform the computation again with topology T_2 in Fig. 1, we will have exactly the same lnL, but b_5+b_6 will be 0 and b_1+b_3 = 0.8239592165 (i.e., the distance between OTUs S1 and S2 is 0.8239592165 as before). This again is perfectly consistent with results from Eq. (1). Topology T_3 in Fig. 1 will lead to the same lnL and the same conclusion with distance between OTUs S1 and S2 being 0.8239592165.

We can also fit the F84 model to the data in Fig. 1a which now has three sites patterns, with the first two sites sharing the first site pattern, sites 3-4 sharing the second site pattern and sites 5-8 sharing the third site pattern. Because the nucleotide frequencies of the four sequences are all equal to 0.25, and because of equal number of transitions and transversions in the sequences so that k = 1, the F84 distance between sequences S1 and S2 is defined by the following likelihood function:

$$
\begin{aligned}
L &= \frac{8!}{2!2!4!} P_s^2 P_v^2 P_{ii}^4 \\
P_{ii} &= \frac{1}{2} e^{-(1+k)D/c} + \frac{1}{4} e^{-D/c} + \frac{1}{4} \\
P_s &= \frac{1}{4} e^{-D/c} - \frac{1}{2} e^{-(1+k)D/c} + \frac{1}{4} \\
P_v &= -\frac{1}{4} e^{-D/c} + \frac{1}{4} \\
c &= 2(\pi_T \pi_C (1 + k/\pi_Y) + \pi_A \pi_G (1 + k/\pi_R) + \pi_R \pi_Y) = 1.25
\end{aligned}
\tag{6}
$$

where D is the F84 distance, P_{ii}, P_s and P_v corresponding to transition probabilities for no change, transition and transversion, respectively. Maximizing L leads to k = 1 (which is expected because we observe the same number of transitions and transversions in the sequences in Fig. 1a) and D = 0.8664339758 which is slightly larger than the JC69 distance.

Applying the pruning algorithm to the four sequences (Fig. 1a) and topology T_1 in Fig. 1, we obtain a final likelihood function that includes only k and (b_1+b_2), i.e., there is no information to estimate b_3, b_4 and b_5+b_6 in topology T_1 in Fig. 1. Maximizing the likelihood function leads to the maximum lnL = -20.79441542, reached when k = 1 and (b_1+b_2) = 0.8664339758. This is exactly the same as the result derived from Eq. (6). If we perform the computation again with topology T_2 in Fig. 1, we will have exactly the same lnL, but b_5+b_6 will be 0 and b_1+b_3 = 0.8664339758 (i.e., the distance between OTUs S1 and S2 is 0.8664339758). This again is perfectly consistent with results from Eq. (6). We can use topology T_3 in Fig. 1 and will again obtain the same lnL and the same conclusion with distance between OTUs S1 and S2 being 0.8664339758.

Note that the application of the F84 model resulted in a small increase in lnL from -21.02998149 with the JC69 model to -20.79441542. This is expected from the sequence data in Fig. 1a which do not conform strictly to the JC69 model. S1 and S2 differ by two transitions and two transversions instead of the 1:2 ratio expected under the JC69 model, so F84 is a more appropriate substitution model than JC69.

The transition/transversion ratio for the DNAML program (R_{DNAML}) is defined [17. p. 18] as

$$R_{DNAML} = \frac{\pi_T\pi_C(1+k/\pi_Y)+\pi_A\pi_G(1+k/\pi_R)}{\pi_R\pi_Y} \tag{7}$$

Given the equal nucleotide frequencies and k = 1, lnL from DNAML is maximized when R_{DNAML} = 1.5, and DNAML outputs lnL = -20.79442 which is the same as shown above. The lnL values are the same for all three topologies. BASEML outputs the same k and lnL. Of course, if one uses DNAML with the default R_{DNAML} of 2, then the three possible topologies will lead to different likelihood values. For this reason, one should not always use default values when running phylogenetic tools. However, misleading phylogenetic results due to misuse of default values should not be attributed to bias in phylogenetic methods.

3 A "Bias" That Is Not True Bias

Suppose we now have the sequence data in Fig. 1b. The four sequences are identical except that S3 and S4 have part of the sequences missing, so there are only two site patterns assuming the JC69 model (with the first shared by the first four sites and the second by the last 32 sites containing unknown nucleotides). These sequences again allow us to have two straightforward expectations. First, the three topologies should have the same lnL. Second, all branches should have length equal to 0 (i.e., $b_i = 0$). Third, the likelihood for each site is simply 0.25, so that the maximum lnL for the entire sequence alignment and for any of the three topologies is

$$\ln L = 4\ln(0.25) + 32\ln(0.25) = -49.906597 \tag{8}$$

which is reached when $b_1 = b_2 = b_3 = b_4 = b_5+b_6 = 0$. One could replace the JC69 model by the F84 model, but the results will be the same because the greater generality of the F84 model relative to the JC69 model is not necessary for the sequence data in Fig. 1b.

Both DNAML and BASEML produce results and conclusions quite different from our expectations when they are used to evaluate the three alternative topologies. First, topology T_1 in Fig. 1 has higher lnL than the other two alternative topologies, and is declared by both DNAML and BASEML as significantly better than the other two alternative topologies. Second, the b_i values listed in the output of DNAML and BASEML are greater than zero and their consequent lnL values are less than the maximum -49.906597 reached when b_i values are all zero.

This "bias" was analytically identified before [Supplemental Materials in 7], and it is not a true bias in the maximum likelihood method. The problem is caused by both DNAML and BASEML not allowing branch lengths to zero during their evaluation of the three alternative topologies. Most likelihood-based phylogenetic programs set a small constant as the lower bound for estimating branch lengths. If we force DNAML and BASEML to evaluate the four-taxon tree with zero branch lengths, they will find lnL to be equal to that in Eq. (8). As soon as we allow branch lengths to be greater

than zero, topology T_1 in Fig. 1 will be favored against the other two alternative topologies by DNAML and BASEML.

The effect is easy to see if we simply set all branch lengths (b_i values) to a small constant C and write down the likelihood functions for the two site patterns (shared by the first four sites and the last 32 sites, respectively) in sequences in Fig. 1b for topologies T_1 and T_2. For T_1, the likelihood functions for the two site patterns ($L_{T1.pattern1}$, $L_{T1.pattern2}$), given the JC69 model, can be obtained by traversing the tree in Fig. 2 and expressed as

$$\begin{aligned} L_{T1.pattern1} &= \frac{1}{4}b^2(b^3+3a^3)+\frac{3}{4}a^2(ab^2+ba^2+2a^3) \\ L_{T1.pattern2} &= \frac{1}{4}b^2+\frac{3}{4}a^2 \\ a &= \frac{1}{4}-\frac{1}{4}e^{-4C/3} \\ b &= \frac{1}{4}+\frac{3}{4}e^{-4C/3} \end{aligned} \tag{9}$$

where all branch lengths are equal to C. Both $L_{T1.pattern1}$ and $L_{T1.Pattern2}$ reach the maximum 0.25 when C = 0 as one would expect.

For topology T_2, the likelihood function for the first site pattern shared by the first four sites is exactly the same as $L_{T1.pattern1}$ in Eq. (9). However, the likelihood function for the second site pattern shared by the 32 "?"-containing sites, is different between topologies T_2 and T_1. For topology T_2, the likelihood function for each of these 32 sites is equal to

$$L_{T2.pattern2} = \frac{1}{4}b(b^2+3a^2)+\frac{3}{4}a(2ab+2a^2) \tag{10}$$

which reaches the maximum 0.25 when C = 0 as one would expect. With the increase in C, $L_{T2.pattern2}$ becomes smaller than $L_{T1.pattern2}$, leading to T_1 preferred over T_2 (or T_3). However, in practical data analysis, the difference should be negligible because the minimum branch length in software is usually set to a value in the order of 0.000001 or smaller. With such a small C, the lnL difference contributed by one site is in the order of 0.000001.

4 True Bias Involving Missing Data Coupled with Among-Site Rate Variation

Suppose now we have sequence data in Fig. 1c, with Gene1 being variable but Gene2, which is missing in S3 and S4, is so conservative as to be invariant. In practice, Gene1 and Gene2 could be different segments within the same gene, e.g., the conserved and variable domains in ribosomal RNAs with no clear boundary between

them. I used this configuration because (1) it has been used before in simulations [7], and (2) it represents a recurring pattern in published supermatrices. Note that the three variable sites at the 5'-end could be diffused at different sites in the data instead of clumping together to be so easily recognizable in real data.

The sequences are intentionally made not to favor any one of the three possible topologies (Fig. 1). For Gene1, the four OTUs are exactly equally divergent from each other given the JC69 and F84 models, i.e., each pair of sequences differ in exactly one transition and two transversions so that no particular topology is favored over the other two. Gene2 is extremely conservative and no substitution has been observed, so it also should not favor any topology over the other two.

With the sequence data in Fig. 1c and topology T_1 in Fig. 1, we can apply the pruning algorithm and the JC69 model to compute the likelihood. There are only three different site patterns with the JC69 model, i.e., sites 1 to 3 share the first site pattern, sites 4 to 7 sharing the second and sites 8 to 39 sharing the third. Maximizing the likelihood will lead to lnL = -83.56464029 which is reached when $b_1 = b_2 = 0.04153005797$, $b_3 = b_4 = 0.3787544804$, and $(b_5+b_6) = 0.3511004094$.

The maximum lnL value for topology T_2 in Fig., 1 is -83.96663731, reached when $b_1 = b_3 = 0.04184900$, $b_2 = b_4 = 0.60765526$, and $(b_5+b_6) = 0.000947018$. The maximum lnL value for topology T_3 is the same as that for T_2 and both are significantly smaller ($p < 0.001$) than that for T_1 (Fig. 1) based on either the Kishino-Hasegawa test and RELL test [16] or Shimodaira & Hasegawa test [18]. DNAML reached exactly the same conclusion, so did BASEML with either the JC69 model or the F84 model. Note that, if the sequence alignment is 100 times as long (which is common in studies with supermatrices), the difference in lnL between T_1 and T_2 would be about 40, which is often greater than the difference between the best and the second best trees in a typical ML reconstruction.

This rejection of topologies T_2 and T_3 in favor of T_1 is not expected from the data in Fig. 1c because each pair of sequences differs by exactly one transition and two transversions. Why is topology T_1 strongly favored by the likelihood method over T_2 and T_3? We can find the answer by making a few observations below.

First, different sites require different branch lengths for maximizing its likelihood. For example, the maximum likelihood for each of the first three sites (Fig. 1c), given the JC69 model, is 0.00390625 ($=0.25^4$) reached when b_1 to b_4 are infinitely large. In contrast, the maximum likelihood for each site from site 4 to site 7 is 0.25 reached when b_1 to b_4 are all zero. Thus, the log-likelihood for the first seven sites (lnL_7), if maximized separately, would be $3*\ln(0.25^4) + 4*\ln(0.25)$, i.e., -22.18070977 for topologies T_1, T_2 and T_3. However, as a compromise between the first three and the next four sites, lnL_7 becomes -32.96754443, reached when $b_1 = b_2 = b_3 = b_4 = 0.3841581410$ and $(b_5+b_6) = 0.1220492271$. This result is applicable to all three topologies. Thus, among-site rate variation itself does not cause phylogenetic bias if it is not lineage-specific, although a previous study [19] suggested that it does based on simulation studies.

Second, the maximum log-likelihood for the 32 sites with missing values in Fig. 1c (lnL_{32}) is also the same among the three topologies, being -44.36141955 when $b_1 = b_2 = 0$ for topology T_1 in Fig. 1a (all other branch lengths are irrelevant for computing lnL_{32} given T_1). For topology T_2 (Fig. 1a) to reach the same maximum lnL_{32}, we need $b_1 = b_3 = (b_5+b_6) = 0$. Similarly, with T_3, we need $b_1 = b_4 = (b_5+b_6) = 0$. Thus, the

branch lengths that maximize $\ln L_{32}$ (i.e., when all branch lengths are zero) are not the same as the branch lengths that maximize $\ln L_7$ (which is maximized when branch lengths are greater than zero, with optimal values specified above), and different topologies impose different constraints on maximizing likelihood.

Third, recall that $\ln L_{32}$ depends only on b_1 and b_2 for topology T_1, but on more branch lengths for T_2 and T_3. With T_1, b_1 and b_2 can be reduced to maximize $\ln L_{32}$ (although not to zero because of the first three variable sites in Fig. 1c). Other branch lengths such as b_3, b_4 and (b_5+b_6) can take optimal values to maximize $\ln L_7$ without affecting $\ln L_{32}$. Because b_1 and b_2 are reduced to maximize $\ln L_{32}$, and consequently deviated substantially from the optimal branch length (= 0.3841581410) for maximizing $\ln L_7$, (b_5+b_6) is increased to 0.3511004094 to compensate. In contrast, $\ln L_{32}$ for topology T_2 depends on b_1, b_3 and (b_5+b_6). Maximization of $\ln L_{32}$ for T_2 can be achieved by reducing b_1, b_3 and (b_5+b_6) and at the same time increasing b_2 and b_4 as a compensation to maximize $\ln L_7$. This explains why the final T_2 tree has relatively short b_1, b_3, both being 0.04184900, and a very small (b_5+b_6), being 0.000947018, but much larger b_2 and b_4, both being 0.60765526.

To recapitulate, maximizing $\ln L_7$ requires $b_1 = b_2 = b_3 = b_4 = 0.3841581410$ and $(b_5+b_6) = 0.1220492271$, and maximizing $\ln L_{32}$ requires $b_1 = b_2 = b_3 = b_4 = (b_5+b_6) = 0$. Obviously, conflicts in maximizing $\ln L_{32}$ and $\ln L_7$ is greater for topology T_2 than for topology T_1, leading to lnL greater for T_1 than for T_2. This result proves the finding by Lemmon et al. [7] reached through sequence simulation, i.e., missing data coupled with among-site rate variation could lead to phylogenetic bias. It should eliminate the doubt expressed on other empirical grounds [6, 11]. One way to eliminate the bias in favor one topology over others is to identify sites with different rates into different partitions. However, in real data, these variable sites may be diffused among conservative sites instead of clumping together as in Fig. 1c to be easily recognizable.

An alternative to partition the sequence alignment is to use a gamma distribution to accommodate rate variation among sites. Unfortunately, parameter estimation (e.g., the shape parameter of the gamma distribution) often depends on topology. Ideally, we should get the same shape parameter regardless of which topology we use, but this is almost never the case. When we get different shape parameters from different topologies, which shape parameter should we trust? If we know that topology T_1 is true, then we would give more credit to the shape parameter obtained with T_1. Alternatively, if we know the true shape parameter, we would trust more the topology that yields a shape parameter that is the same as the true parameter than other topologies that generate a shape parameter that is far from the true value. Such a chicken-egg problem lands us in an awkward dilemma.

Note that the first four sites in Fig. 1c are equivalent to a stretch of the alignment that has undergone substitution saturation. While phylogenetic information will be eroded by substitution saturation and tests have been developed to assess such substitution saturation [20, 21], it is perhaps the first time to link substitution saturation directly to phylogenetic bias in the context of missing data. Also note that, although sequences in Fig. 1c is biased in favor of grouping S_1 and S_2 together, one can easily envision scenarios in which S_1 and S_2 would repulse each other, e.g., when the last 32 sites in Fig. 1c are far more variable than the first seven sites. Thus, the direction of the bias cannot be predicted before data analysis.

Lemmon et al. [7] speculated that the bias they observed from simulated sequences with missing data may be associated with model misspecification. While there is possibility for such an association, the results I have presented show that the bias can be entirely independent of model misspecification.

The bias associated with missing data and rate heterogeneity among sites has been noted for a long time. For example, the 18S rRNA sequences contain the variable and conservative regions. Missing a variable region or a conservative region by a subset of sequences leads to distortion of phylogenetic signals and wrong phylogenetic trees [22, 23]. Dramatic rate heterogeneity among the three codon positions [24, 25], or among genes located in different DNA strands [27, 28] have long been noted. As among-site rate variation is not only common in molecular sequence data but also a known source of phylogenetic bias [19], one should be cautious to compile such data with missing data configuration similar to that in Fig. 1c. As a precaution against such bias, some computer programs, e.g., DAMBE [29], deletes sites containing missing data before likelihood analysis.

One may not consider this as a serious problem because, among all those compiled supermatrices in recent publications [e.g., 1, 2, 3], closely related species tend to share genes (or lack of genes). If we take the data in Fig. 1b as a simple caricature of the supermatrices, S1 and S2 are more likely to be closely related to each other, so are S3 and S4, in real data compilations. This means that the bias above caused by missing data will tend to help recover the true topology or increase the bootstrap support of some true subtrees. This may well have contributed to the increased bootstrap values documented before by Cho et al. [11] who then have argued for the supermatrix approach based on increased bootstrap values for certain taxa. Such an argument is flawed. We may recall an analogous case in the maximum parsimony method, with the inconsistency caused by long-branch attraction. Closely related species generally are more likely to share long branches than remote species, so long-branch attraction could seem a good thing because the bias it causes may lead to more efficient recovering of the true tree or increase bootstrap support for some true subtrees. However, such increased efficiency in recovering the true tree or increased bootstrap support for some true subtrees is purchased with illegal phylogenetic currency and should always be discouraged. In statistical estimation, a bias is a bias and is always undesirable because it often renders results unpredictable. It is fortunate that there has been only one case in which a phylogenetic approach is justified by its bias/inconsistency [30].

In summary, many supermatrices laden with missing sequences have been compiled in recent years while few studies have been carried out on the potential statistical bias that such data may cause. While the likelihood method handles missing data in a sensible way, its implementation may not achieve sufficient precision and may cause phylogenetic bias induced by missing data. In particular, lumping genes with different evolutionary rates runs a high risk of distorting phylogenetic signals and likelihood values and should be strongly discouraged.

Acknowledgements. This study is supported by Discovery Grant from Natural Science and Engineering Research Council (NSERC) of Canada. I thank B. Foley and X. Sun for motivating me to write the paper, and D. Baurain, S. Aris-Brosou, B. Golding, and A. RoyChoudhury for discussion and comments.

References

1. Hackett, S.J., Kimball, R.T., Reddy, S., Bowie, R.C., Braun, E.L., Braun, M.J., Chojnowski, J.L., Cox, W.A., Han, K.L., Harshman, J., Huddleston, C.J., Marks, B.D., Miglia, K.J., Moore, W.S., Sheldon, F.H., Steadman, D.W., Witt, C.C., Yuri, T.: A phylogenomic study of birds reveals their evolutionary history. Science 320, 1763–1768 (2008)
2. Perelman, P., Johnson, W.E., Roos, C., Seuanez, H.N., Horvath, J.E., Moreira, M.A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M.P., Silva, A., O'Brien, S.J., Pecon-Slattery, J.: A molecular phylogeny of living primates. PLoS Genet. 7, e1001342 (2011)
3. Regier, J.C., Shultz, J.W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., Martin, J.W., Cunningham, C.W.: Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463, 1079–1083 (2010)
4. Regier, J.C., Shultz, J.W., Ganley, A.R., Hussey, A., Shi, D., Ball, B., Zwick, A., Stajich, J.E., Cummings, M.P., Martin, J.W., Cunningham, C.W.: Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57, 920–938 (2008)
5. Sanderson, M.J., Ane, C., Eulenstein, O., Fernandez-Baca, D., Kim, J., McMahon, M.M., Piaggio-Talice, R.: Fragmentation of large data sets in phylogenetic analysis. In: Gascuel, O., Steel, M. (eds.) Reconstructing Evolution: New Mathematical and Computational Advances, pp. 199–216. Oxford University Press, Oxford (2007)
6. Wiens, J.J., Tiu, J.: Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. PLoS One 7, e42925 (2012)
7. Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M.: The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145 (2009)
8. Wiens, J.J.: Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538 (2003)
9. Roure, B., Baurain, D., Philippe, H.: Impact of Missing Data on Phylogenies Inferred from Empirical Phylogenomic Data Sets. Mol. Biol. Evol. 30, 197–214 (2013)
10. Rubin, B.E., Ree, R.H., Moreau, C.S.: Inferring phylogenies from RAD sequence data. PLoS One 7, e33394 (2012)
11. Cho, S., Zwick, A., Regier, J.C., Mitter, C., Cummings, M.P., Yao, J., Du, Z., Zhao, H., Kawahara, A.Y., Weller, S., Davis, D.R., Baixeras, J., Brown, J.W., Parr, C.: Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol. 60, 782–796 (2011)
12. Felsenstein, J.: Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22, 240–249 (1973)
13. Felsenstein, J.: Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376 (1981)
14. Felsenstein, J.: Inferring phylogenies. Sinauer, Sunderland (2004)
15. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. In: Munro, H.N. (ed.) Mammalian Protein Metabolism, pp. 21–123. Academic Press, New York (1969)
16. Kishino, H., Hasegawa, M.: 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 (1989)
17. Yang, Z.: Computational molecular evolution. Oxford University Press, Oxford (2006)
18. Shimodaira, H., Hasegawa, M.: Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Mol. Biol. Evol. 16, 1114–1116 (1999)

19. Kuhner, M.K., Felsenstein, J.: A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468 (1994)
20. Xia, X., Lemey, P.: Assessing substitution saturation with DAMBE. In: Lemey, P., Salemi, M., Vandamme, A.M. (eds.) The Phylogenetic Handbook, pp. 615–630. Cambridge University Press, Cambridge (2009)
21. Xia, X.H., Xie, Z., Salemi, M., Chen, L., Wang, Y.: An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26, 1–7 (2003)
22. Van de Peer, Y., Neefs, J.M., De Rijk, P., De Wachter, R.: Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. J. Mol. Evol. 37, 221–232 (1993)
23. Xia, X.H., Xie, Z., Kjer, K.M.: 18S ribosomal RNA and tetrapod phylogeny. Syst. Biol. 52, 283–295 (2003)
24. Xia, X., Hafner, M.S., Sudman, P.D.: On transition bias in mitochondrial genes of pocket gophers. J. Mol. Evol. 43, 32–40 (1996)
25. Xia, X.: The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. Mol. Biol. Evol. 15, 336–344 (1998)
26. Marin, A., Xia, X.: GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. J. Theor. Biol. 253, 508–513 (2008)
27. Xia, X.: DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. Current Genomics 13, 16–27 (2012)
28. Xia, X.: DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol. Biol. Evol. 30, 1720–1728 (2013)
29. Siddall, M.E.: Success of Parsimony in the Four-Taxon Case: Long-Branch Repulsion by Likelihood in the Farris Zone. Cladistics 14, 209–220 (1998)