Evolutionary Bioinformatics

OPEN ACCESS
Full open access to this and thousands of other papers at http://www.la-press.com.

ORIGINAL RESEARCH

RY-Coding and Non-Homogeneous Models Can Ameliorate the Maximum-Likelihood Inferences From Nucleotide Sequence Data with Parallel Compositional Heterogeneity

Sohta A. Ishikawa[1], Yuji Inagaki[1,2] and Tetsuo Hashimoto[1,2]

[1]Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8572, Japan.
[2]Center for Computational Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8577, Japan.
Corresponding author email: yuji@ccs.tsukuba.ac.jp

Abstract: In phylogenetic analyses of nucleotide sequences, 'homogeneous' substitution models, which assume the stationarity of base composition across a tree, are widely used, albeit individual sequences may bear distinctive base frequencies. In the worst-case scenario, a homogeneous model-based analysis can yield an artifactual union of two distantly related sequences that achieved similar base frequencies in parallel. Such potential difficulty can be countered by two approaches, 'RY-coding' and 'non-homogeneous' models. The former approach converts four bases into purine and pyrimidine to normalize base frequencies across a tree, while the heterogeneity in base frequency is explicitly incorporated in the latter approach. The two approaches have been applied to real-world sequence data; however, their basic properties have not been fully examined by pioneering simulation studies. Here, we assessed the performances of the maximum-likelihood analyses incorporating RY-coding and a non-homogeneous model (RY-coding and non-homogeneous analyses) on simulated data with parallel convergence to similar base composition. Both RY-coding and non-homogeneous analyses showed superior performances compared with homogeneous model-based analyses. Curiously, the performance of RY-coding analysis appeared to be significantly affected by a setting of the substitution process for sequence simulation relative to that of non-homogeneous analysis. The performance of a non-homogeneous analysis was also validated by analyzing a real-world sequence data set with significant base heterogeneity.

Keywords: RY-coding, non-homogeneous model, model misspecification, long-branch attraction, compositional heterogeneity

Evolutionary Bioinformatics 2012:8 357–371

doi: 10.4137/EBO.S9017

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article. Unrestricted non-commercial use is permitted provided the original work is properly cited.

Introduction

In molecular phylogenetic analyses, two distantly related, rapidly evolving (long-branch) sequences often erroneously group together owing to long-branch attraction (LBA) artifacts, which have been recognized as one of the major sources of phylogenetic artifacts.[1] Pioneering studies based on simulated data have shown that the susceptibility to LBA artifacts differs amongst tree reconstruction methods: distance matrix (DM)-based and maximum-parsimony methods are sensitive to, but the maximum-likelihood (ML) method is in theory robust against LBA artifacts.[2,3] This ideal property of the ML method, however, collapses under conditions such as 'model misspecification,' where the substitution model applied to a phylogenetic analysis does not sufficiently describe the evolutionary process that generated the sequence data. As the precise evolutionary process underlying real-world sequences is difficult to know, there is always a risk of a critical aspect (or aspects) in sequence evolution being overlooked by phylogenetic analysis with a particular substitution model. Therefore, depending on the degree of model misspecification, the ML inference can suffer from severe LBA artifacts.

As base composition varies amongst genomes or even within a single genome, compositional heterogeneity is likely ubiquitous in nucleotide (nt) alignments for phylogenetic analyses. However, widely used 'homogeneous' models for sequence substitutions assume the homogeneity of base composition across a tree, estimating the averaged base frequencies from the entire alignment. Thus, analyzing nt data bearing extreme compositional heterogeneity under homogeneous model conditions introduces significant model misspecification to tree reconstruction, resulting in severe LBA artifacts. Indeed, LBA artifacts attributed to ignoring compositional heterogeneity in nt alignments, as well as amino acid alignments, have been documented in the analyses of both real-world[4–8] and simulated data.[9–12]

In DM-based analyses, LogDet transformation,[13,14] which can take compositional heterogeneity into account, has been widely used for the analyses of alignments bearing significant compositional heterogeneity. Past simulation studies have shown that LogDet distance method outperforms DM-based methods with 'homogeneous' models in recovering the correct tree from alignments with a high degree of compositional heterogeneity.[11,12] Unfortunately, LogDet distance method is not always a practical solution for countering artifacts stemming from compositional heterogeneity in analyses of real-world sequences. Sequences with extreme compositional bias often are more rapidly evolving than other homologues with unbiased composition.[15,16] As LogDet distance method is apparently DM-based, studies based on both real-world and simulation data showed this method to be susceptible to typical LBA artifacts, yielding the artifactual union of rapidly evolving sequences.[10,11,17,18]

The degree of compositional heterogeneity in sequence data can be reduced by character recoding. In nt alignments, the variation of AT (or GC) content across a tree can be efficiently diminished by recoding four characters, A, C, G, and T, into purine (R; A or G) or pyrimidine (Y; C or T).[19,20] This 'RY-coding,' coupled with an ML method, is believed to prevent the putative artifact stemming from the heterogeneity of AT content across a tree and ameliorate the accuracy of phylogenetic inferences. Nevertheless, this procedure cannot erase compositional heterogeneity among any sequences except those with the ratio of A plus G to C plus T being roughly 1, suggesting that a certain degree of compositional heterogeneity remains in the recoded data. As the recoded alignments are usually analyzed by the ML method with the homogeneous substitution model proposed by Cavender and Felsenstein,[21] it is naïve to assume that the ML inferences from the recoded alignments are liberated from the phylogenetic artifacts from compositional heterogeneity. Finally, the recoding procedure may discard informative transition substitutions (A↔G or T↔C) in the original alignments. Importantly, the efficacy and limitation of RY-coding remain uncertain, as no simulation study assessing the concerns mentioned above is available.

'Non-homogeneous' models can explicitly take compositional heterogeneity across a tree into account.[22] A study based on simulated nt data with biased base composition evidently showed that the accuracy of a DM-based method was improved by a non-homogeneous model.[9] An analysis with a non-homogeneous model requires no character recoding in an alignment, being free from the potential issues associated with RY-coding discussed above. Furthermore, the ML method with a non-homogeneous model is anticipated to be much more robust against typical LBA artifacts than any DM-based methods with or without LogDet transformation. However, to our knowledge, the robustness of ML inferences under non-homogeneous model conditions has not yet been examined in detail by analyzing simulated data.

We here present the results from the de facto first simulation study assessing the performance of an ML method incorporating RY-coding and that with a non-homogeneous model. Simulated nt sequence data bearing various degrees of compositional heterogeneity were subjected to the two types of ML analyses. Our study indicated that the ML analyses incorporating RY-coding and a non-homogeneous model (henceforth designated as RY-coding and nonhomogeneous analyses, respectively) were more robust against the LBA artifact stemming from compositional heterogeneity than the ML analysis with a homogeneous model (henceforth designated as homogeneous analysis). Additionally, a real-world data set with a difference in AT content across a tree (ΔAT%) of 25% was subjected to non-homogeneous analysis. Significantly, as anticipated from the results from our simulated study, non-homogeneous analysis successfully suppressed the artifactual tree topology recovered in a homogeneous analysis, which cannot account for a large ΔAT%.

Our closed investigation, however, revealed the potential pitfalls of both RY-coding and non-homogeneous analyses. The performance of RY-coding analysis appeared to be largely affected by the substitution process used for sequence simulation. Likewise, the inference from non-homogeneous analysis can be significantly misled when the substitution process was incorrectly modeled. Practically, we recommend cross-checking the inference from the ML method with RY-coding by that from the ML method with a non-homogeneous model, which can sufficiently account for the substitution process in the alignment of interest.

Materials and Methods

Data simulation

Nucleotide sequence data were generated by Monte Carlo simulation, using indel-Seq-Gen Version 2.0,[23] based on a 4-taxon model tree described below (Fig. 1A). For each data point, we simulated 500 replicates. The simulated data were varied from 500, 1000, 2500, and 5000 nt positions in size. The lengths for the central branch and two terminal branches leading to Taxa 1 and 2 were set to 0.025, and the lengths of the terminal branches leading to Taxa 3 and 4 were set to 0.800 (*a* and *b* in Fig. 1A). These specific branch lengths were determined on the basis of preliminary analyses of sequence data simulated over 4-taxon model trees with 1600 combinations of branch lengths *a* and *b* (with *a* ranging from 0.0125 to 0.5000, and *b* ranging from 0.5000 to 1.0000; see Fig. S1). For each simulation, the ancestral sequence was randomly generated at the root (R in Fig. 1A), and each tip sequence was then simulated according to the given branch lengths. The substitution process was modeled with the HKY model,[24] incorporating rate heterogeneity across sites approximated by a discrete gamma (Γ) distribution with four categories (HKY + Γ model). The κ parameter for transition/transversion (Ts/Tv)

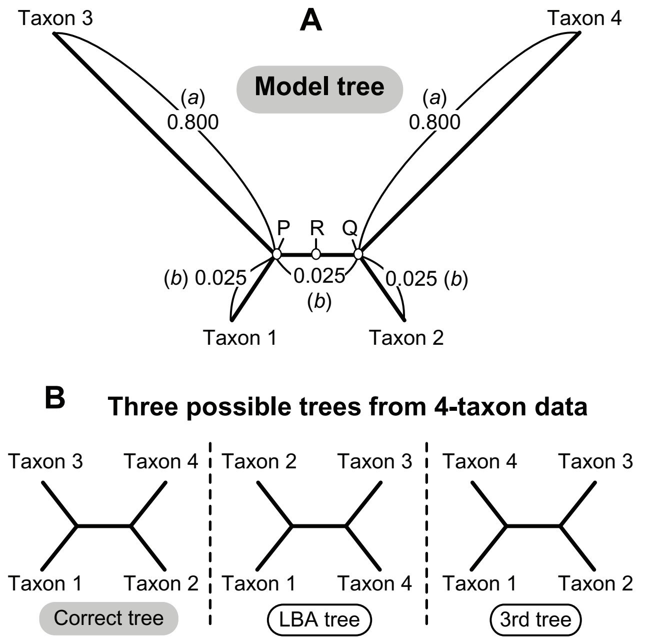


Figure 1. Four-taxon trees considered in this study. (**A**) A model tree for sequence simulation. The lengths of the terminal branches leading to Taxa 3 and 4 were set as 0.800, while those of the rest of branches in the tree were set as 0.025. In this figure, the branch lengths were not correctly scaled for readers' convenience. Firstly, random sequences with AT content of ~50% were generated at the root (R). Subsequently, Taxa 1–4 sequences were simulated based on the given 'root' sequence, branch lengths, and model parameters. The parameters for discrete gamma distribution and transition/transversion ratio were fixed across a tree. The frequencies for A, C, G, and T were set equal from the root to the terminal branches leading to Taxa 1 and 2, while unequal frequencies for the four bases were applied to the terminal branches leading to Taxa 3 and 4. The parameters for the base frequencies applied to the branches leading to Taxa 3 and 4 are shown in Table 1. (**B**) Possible tree topologies from the 4-taxon simiulated data. Branch lengths are not scaled.

ratio and the shape parameter α for a Γ distribution were set to 2.0 and 0.8, according to Galtier and Gouy.[9] We additionally simulated data with smaller κ values, 0.2, 0.5, 1.0, and 1.5, to evaluate how the setting of Ts/Tv ratio in sequence simulation affects the performnce of the ML analyses.

For the simulation from the root to Taxa 1 and 2, the frequencies of A, C, G, and T were set equal (ie, the AT content is supposed to be ~50%). On the other hand, Taxa 3 and 4 sequences were designed to be AT-rich by changing the parameters for base frequency at the node uniting Taxa 1 and 3, and that uniting Taxa 2 and 4 (P and Q, respectively, in Fig. 1A). The above procedure enabled us to simulate slowly evolving sequences for Taxa 1 and 2 with an AT content of ~50%, and rapidly evolving, AT-rich sequences for Taxa 3 and 4. We analyzed the simulated data with 11 variations of ΔAT% calculated between slowly evolving Taxa 1 and 2, and rapidly evolving Taxa 3 and 4. The frequencies of A and T and those of C and G were set equal unless we specifically mention. We provide the settings for base frequency in the data simulation, and the average AT% achieved in the resultant simulated data in Table 1.

Data analyses

We ran three different ML analyses in the present study. First, the simulated data (comprising four nt characters) were subjected to the ML analysis with the HKY + Γ model. The Ts/Tv ratio and shape parameter α for a Γ distribution were fixed to those used in the data simulation (κ = 2.0–0.2, and α = 0.8), but base frequencies were estimated from the entire data. We also analyzed the simulated data recoded by RY-coding.[19,20] The recoded data (comprising binary characters) were then analyzed with the model of Cavender and Felsenstein[21] for two-state characters incorporating rate heterogeneity across sites approximated by a discrete Γ distribution (CF + Γ model). All model parameters for the second ML analysis were estimated from the data. The substitution models used in the first and second ML analyses are homogeneous as they assume the stationarity of substitution process. We used PAUP* 4.0b,[25] for the ML analyses with the two homogeneous models.

Finally, we subjected the simulated data to the third ML analysis with a non-homogeneous model proposed by Galtier and Gouy[22] incorporating rate heterogeneity across sites approximated by a discrete Γ distribution (GG98 + Γ model) implemented in NHML 3.0.[26] In this non-homogeneous model, the parameters for Ts/Tv ratio and the Γ distribution were estimated from the entire data, but the parameter for AT content was allowed to vary in a branch-by-branch fashion. We exhaustively searched for the ML tree by the eval_nh program packaged in NHML.[26] In addition, a subset of simulated data was analyzed with a second non-homogeneous model, which is identical to the HKY + Γ model but allows base frequencies to vary across a tree (nhHKY + Γ model).

Table 1. Settings for the base frequencies applied to the terminal branches leading to Taxa 3 and 4, and the average AT content (AT%) in the resultant Taxa 3 and 4 sequences.

Ts/Tv ratio (κ) = 2.0			**Ts/Tv ratio (κ) = 0.2**		
Settings of base frequencies in data simulation (%)		**AT% achieved in the simulated data (%) (mean ± 2*SD)**	**Settings of base frequencies in data simulation (%)**		**AT% achieved in the simulated data (%) (mean ± 2*SD)**
A and T	**G and C**		**A and T**	**G and C**	
25.0	25.0	50.0 ± 2.5	25.0	25.0	50.0 ± 2.7
26.5	23.5	51.7 ± 2.5	27.0	23.0	51.9 ± 2.6
28.0	22.0	53.4 ± 2.6	29.0	21.0	53.8 ± 2.6
29.5	20.5	55.1 ± 2.5	31.0	19.0	55.8 ± 2.5
31.0	19.0	56.8 ± 2.6	33.0	17.0	57.7 ± 2.7
32.5	17.5	58.6 ± 2.4	35.0	15.0	59.7 ± 2.6
34.0	16.0	60.5 ± 2.5	37.0	13.0	61.9 ± 2.7
35.5	14.5	62.4 ± 2.4	39.0	11.0	64.1 ± 2.6
37.0	13.0	64.5 ± 2.4	41.0	9.0	66.5 ± 2.6
38.5	11.5	66.7 ± 2.6	43.0	7.0	69.4 ± 2.4
40.0	10.0	69.2 ± 2.4	45.0	5.0	72.7 ± 2.4

Analyses of a real-world data set

We retrieved the gene sequences encoding four ribosomal proteins (L14, L16, S3, and S11) and β subunit of RNA polymerase encoded in 9 red algal or red alga-derived plastids, 17 green algal or green alga-derived plastids, and five residual plastids in apicomplexan parasites (so-called apicoplasts) from GenBank database. For each gene, nt sequences were carefully aligned by referring their putative amino acid sequences. After the exclusion of unambiguously aligned positions, the five single-gene alignments were concatenated into a '5-gene' alignment containing 31 taxa with 2226 nt positions. Of note, the AT contents of the apicoplast sequences are generally higher than other plastid sequences and produce a wide range of ΔAT% in the 5-gene alignment (Fig. 2A).

We firstly conducted the ML analyses of the original 5-gene alignment and 100 bootstrap replicates with the HKY + Γ model by PhyML3.0.[27] All parameters were estimated from the entire data. The ML analysis placed the clade of the five apicoplasts within the green algal/green alga-derived plastids (Fig. 2A), implying that apicoplasts were established through secondary endosymbiosis of a green alga (henceforth here designated as the 'green origin' of apicoplasts; Fig. 2B). Nevertheless, this result is contradictory to the widely accepted the 'red origin' of apicoplasts,[28] which regards apicoplasts as a residual endosymbiotic red alga (Fig. 2B). We conjectured that the tree topology representing the green origin of apicoplasts was attributed to the homogeneous (HKY) model ignoring the heterogeneity of AT content in the 5-gene alignment (see the bar graph in Fig. 2A). If the above conjecture is true, we can anticipate that the ML analysis with a non-homogeneous model may suppress the phylogenetic artifact and recover a tree topology representing the red origin of apicoplasts.

On the basis of the results presented in Figure 2A, it is difficult to evaluate how the ML tree representing the green origin of apicoplasts was superior to the trees representing the alternative hypothesis. To evaluate the two competing hypotheses for the origin of apicoplasts, we prepared alternative trees by modifying the ML tree shown in Figure 2A. The apicoplast clade was regrafted to (i) seven terminal branches showing no apparent affinity to any other sequences, or (ii) seven branches which are basal to the robustly supported clades (highlighted by dots in Fig. 2A). Subsequently, the log-likelihoods (lnLs) of the 14 alternative trees were compared with that of the ML tree (see below).

The lnLs of the ML and 14 alternative trees were calculated with the HKY + Γ (homogeneous) model by PhyML3.0.[27] The same calculation was repeated with a non-homogeneous (GG98 + Γ) model by eval_nh.[26] The root position was fixed in the second comparison with the GG98 + Γ model (highlighted by a diamond in Fig. 2A). Branch lengths were optimized during the lnL calculation. Model parameters were estimated from the entire alignment.

Results

Impact of compositional heterogeneity—HKY analysis

The HKY model assumes the stationarity of the substitution process (ie, homogeneous), and ΔAT% in the simulated data cannot be adequately accounted for. Henceforth here, we designate the HKY model-based ML analysis as 'HKY analysis.' On the basis of Jermiin et al[12] and Ho and Jermiin,[11] we expected that 'LBA' tree (center in Fig. 1B), in which rapidly evolving Taxa 3 and 4 erroneously grouped together, was preferentially recovered in HKY analysis of the data bearing large ΔAT%.

Indeed, in the analysis of 1000 nt-long data simulated with κ = 2.0 (henceforth designated as 'κ_2.0 data'), the recovery rate of the correct tree (left in Fig. 1B) gradually decreased along with the increment of ΔAT% (black circles in Fig. 3A). On the other hand, LBA tree was dominantly yielded in the analyses of the data with high ΔAT% (Fig. S2). A similar but clearer trend for the success rate (as well as the recovery rate for LBA tree) was observed in the analysis of the data simulated with κ = 0.2 (henceforth designated as 'κ_0.2 data'; black circles in Fig. 3B). These results evidently suggest that HKY analysis, particularly when the data bear large ΔAT%, becomes highly susceptible to the LBA artifact stemming from compositional heterogeneity.

We additionally tested how the performance of HKY analysis was affected by the Ts/Tv ratio in data simulation. Five sets of 1000 nt-long data bearing ΔAT = ~20% were simulated with different κ values, 0.2, 0.5, 1.0, 1.5, and 2.0, and subjected to HKY analysis. As shown in Figure 3C, the analysis of κ_2.0 data yielded the highest success rates (~30%),

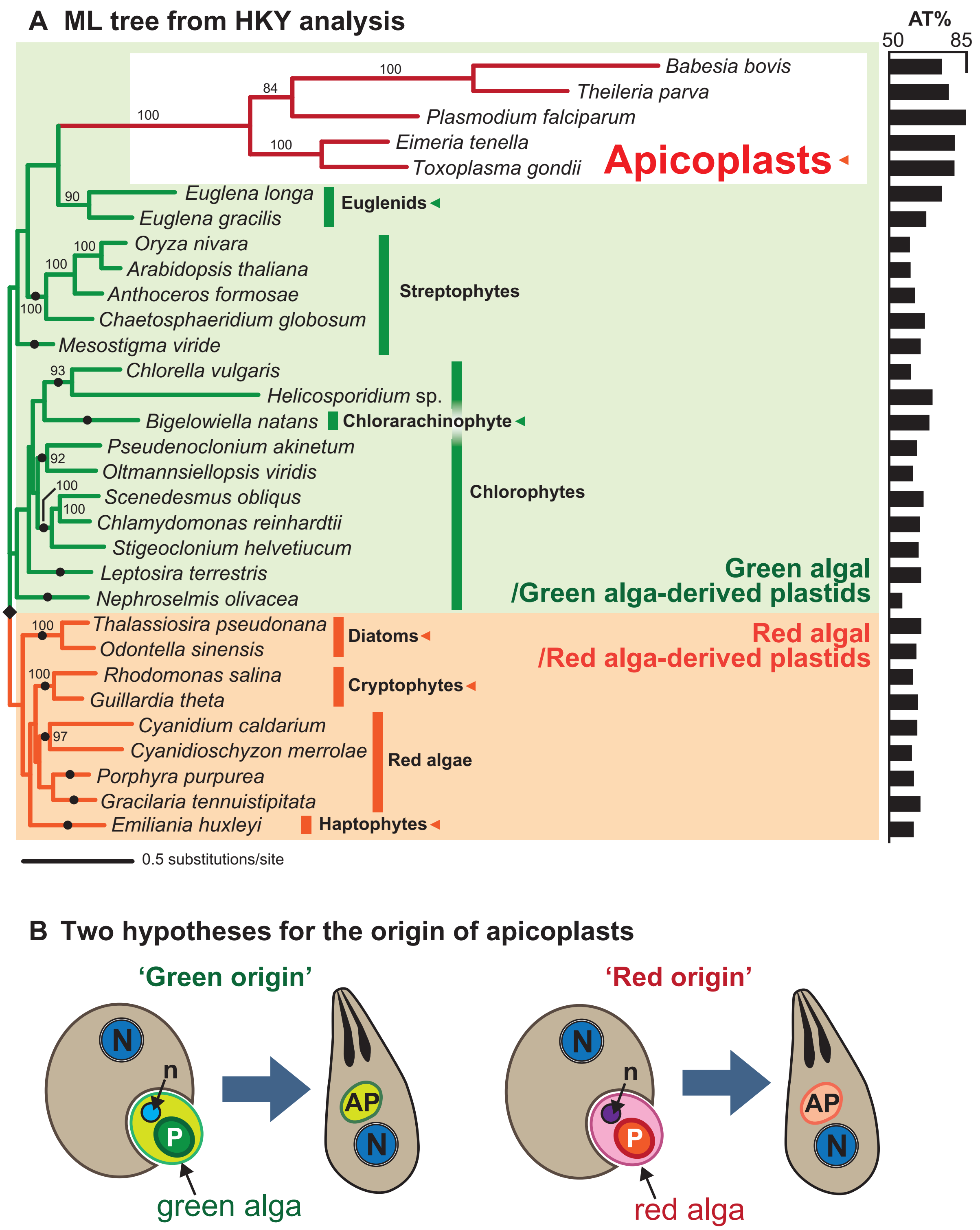


Figure 2. The origin of apicoplasts. (**A**) Maximum-likelihood tree inferred from the 5-gene alignment with the HKY + Γ (homogeneous) model. The subtree for red algal/red alga-derived plastids is in orange, while that for green algal/green alga-derived plastids is in green. The subtree for the residual plastids in apicomplexan parasites (apicoplasts) is in red. Green alga- and red alga-derived plastids are highlighted by green and orange arrowheads, respectively. In this topology, the apicoplast clade is placed within green algal/green alga-derived plastids, representing the 'green origin' of apicoplasts. For each taxon, the AT content (AT%) is shown on the right side. Bootstrap proportions larger than 50% are shown for the nodes. (**B**) Hypothetical origin of apicoplasts. The scheme on the left represents the 'green origin' of apicoplasts—apicoplasts are the descendants of an endosymbiotic green alga. On the other hand, the 'red origin' of apicoplasts schematically shown on the right assumes that apicoplasts were derived from an endosymbiotic red alga.
Notes: The nucleus of the endosymbiotic alga (n) has disappeared in modern apicomplexan cells.
Abbreviations: N, host nucleus; n, endosymbiotic algal nucleus; P, plastid; Ap, apicoplast.

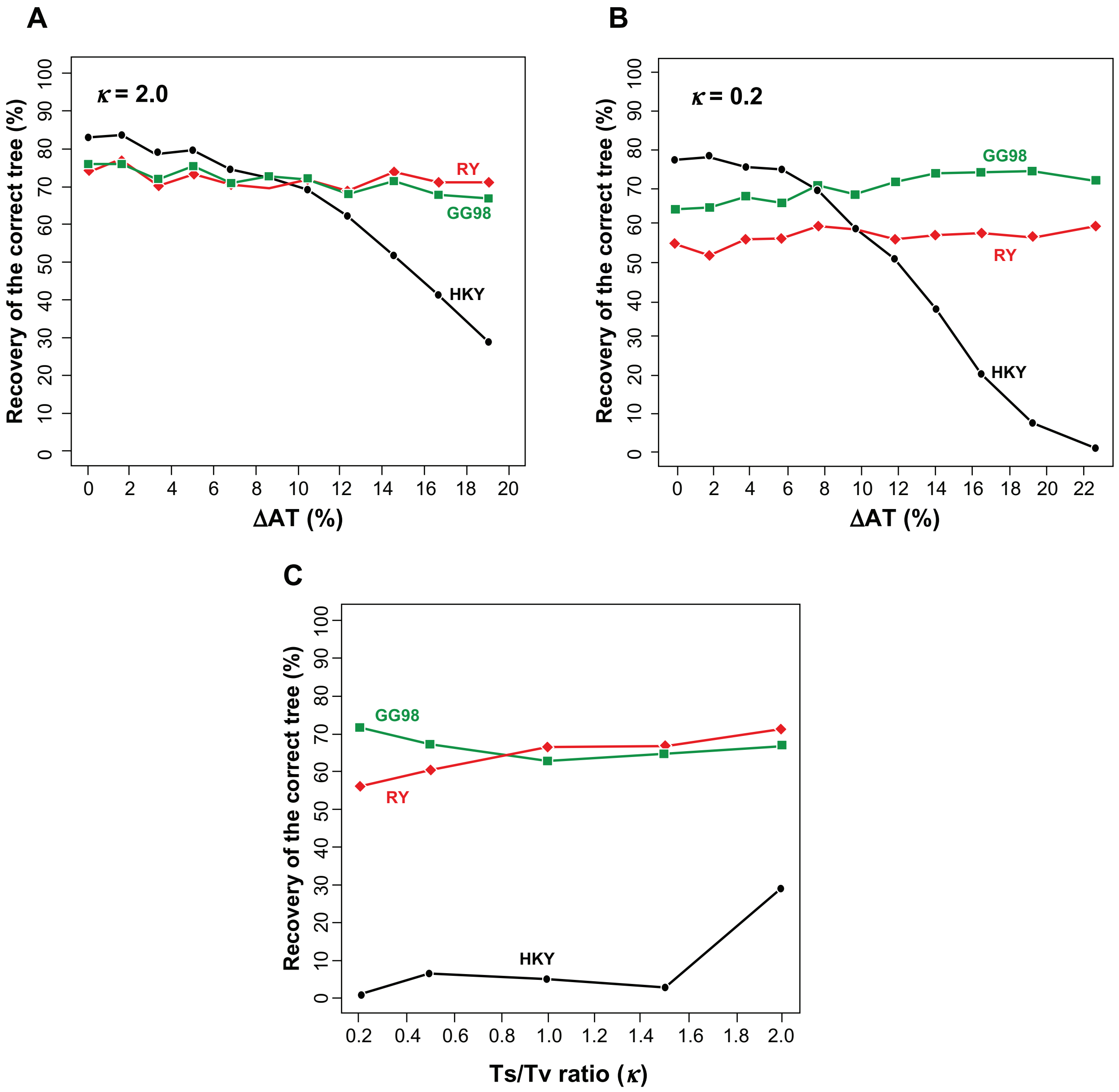


Figure 3. Impacts of the difference in AT content across a tree (ΔAT) and transition/transversion ratio (κ) on the recovery rate of the correct tree. (**A**) Analysis of 4-taxon data simulated with κ = 2.0. We prepared 11 sets of 500 replicates of 1000 nt-long sequence data simulated with different ΔAT%. The simulated data were subjected to the maximum-likelihood analyses with the HKY + Γ model (HKY; black circles) and the GG98 + Γ model (GG98; green squares). We also recoded the simulated data (comprising four nt characters, A, C, G, and T) into binary characters, purine (R; A or G) and pyrimidine (Y; T or C), and then subjected to the ML analysis with the CF + Γ model (RY; red diamonds). (**B**) Analysis of 4-taxon data simulated with κ = 0.2. The details are same as described in (**A**), except κ was set as 0.2. (**C**) Analysis of 4-taxon data simulated with five different κ values. We prepared five sets of 500 replicates of 1000 nt-long sequence data simulated with a fixed ΔAT of ~20%, but κ of 0.2, 0.5, 1.0, 1.5, or 2.0.

while the correct tree was recovered at less than 10% in the analyses of the data simulated with κ < 2.0.

Impact of compositional heterogeneity—RY-coding analysis

RY-coding has been widely used for the analyses of real-world nt data bearing base compositional bias.[19,29] However, there is a (potentially large) room for argument on whether this procedure can truly help in reconstructing the correct tree. In this study, both κ_2.0 and κ_0.2 data series bearing ΔAT of 0–20% were subjected to RY-coding analysis.

We firstly checked whether the recoding procedure erased the compositional heterogeneity simulated in κ_2.0 and κ_0.2 data. Regardless of the degree of ΔAT% in the original data or the setting for Ts/Tv ratio in data simulation, the ratio of R to Y was almost equal (data not shown). As almost no compositional heterogeneity existed, the correct tree was stably recovered in the homogeneous (CF model-based) analyses of the recoded κ_2.0 and κ_0.2 data at 69%–77% and 53%–60%, respectively (red diamonds in Fig. 3A and B). The recovery of LBA tree was less than 18% and 29% in the analyses of the recoded κ_2.0 and κ_0.2 data series, respectively (Fig. S2). The success rate of RY-coding analysis remained high irrespective of Ts/Tv ratio (56%–70%; red diamonds in Fig. 3C), compared with that of HKY analysis. We successfully provide the first simulation results that indicate that RY-coding largely improved the phylogenetic inferences of sequence data with compositional heterogeneity.

Impact of compositional heterogeneity—GG98 analysis

The non-homogeneous GG98 model proposed by Galtier and Gouy in 1998[22] allows different AT% on different branches. The GG98 model has been applied for the ML analyses of real-world sequence data, and successfully displayed the robustness against systematic artifacts originating from compositional heterogeneity.[30] A simulation study by Galtier and Gouy in 1995[9] showed that the GG98 model drastically improved the accuracy of a DM-based analysis, but the performance of GG98 model-based ML analysis (henceforth here designated 'GG98' analysis) has not been tested. In the present study, we examined how the GG98 model can improve the ML inference from sequence data with large ΔAT%.

Regardless of ΔAT%, the correct tree was recovered at 67%–76% in the analysis of κ_2.0 data series (green squares in Fig. 3A), while the recovery of LBA tree was suppressed (<23%; Fig. S2). In GG98 analysis of κ_0.2 data series, ΔAT% had little impact on the success rate (63%–72%; green squares in Fig. 3B). The same analysis was repeated on the 1000 nt-long data simulated with the five different Ts/Tv ratios (ΔAT was set as ~20%), but the success rates stayed at 63%–72% (Fig. 3C). These are the first simulation results indicating that the parallel shifts of base frequency in sequence data could be tolerated in non-homogeneous model-based ML analysis.

Impact of data size

We simulated 500, 1000, 2500, and 5000 nt-long data with ΔAT = ~20%, and these data were subsequently subjected to HKY, RY-coding, and GG98 analyses. The data simulated with the largest and smallest κ values, 2.0 and 0.2, were considered in these analyses. The success rates obtained from the three ML analyses were plotted in Figure 4A and B. Regardless of κ parameter, the success rate of HKY analysis appeared to be negatively correlated with data size (black circles in Fig. 4A and B). The analyses of the largest κ_2.0 and κ_0.2 data (ie, 5000 nt-long) marked the lowest success rates, 14% and 0%, respectively. The magnitude of the LBA artifact stemming from compositional heterogeneity was apparently enhanced by increment of data size.

The success rates of RY-coding analysis positively correlated with data size, and this trend was independent from the setting of κ parameter (red diamonds in Fig. 4A and B). The highest success rates were 96% and 84% in the analyses of the largest κ_2.0 data and the largest κ_0.2 data, respectively. In GG98 analyses of the two data simulated with two different κ values, the success rates were improved by increment of data size (up to 95% and 98%, respectively; green squares in Fig. 4A and B). These plots clearly suggest that data size can further enhance the performances of RY-coding and GG98 analyses against the LBA artifact from compositional heterogeneity in the data.

GG98 analysis versus RY-coding analysis

Both RY-coding and GG98 analyses were robust against ΔAT% in the simulated data (Fig. 3A and B), and their success rates displayed positive correlation with data size (Fig. 4A and B). However, the success rates from GG98 analyses of κ_0.2 data series were constantly greater than the corresponding values from RY-coding analyses (Fig. 4B). We statistically compared the success rates of 500 simulation trials from RY-coding and GG98 analyses for 500, 1000, 2500, and 5000 nt-long κ_0.2 data by Pearson's chi-square test. In all the comparisons, the null hypothesis of the success rate being the same between the RY-coding and GG98 analyses was rejected with extremely small *P* values ($P = 5.2 \times 10^{-6}$–2.2×10^{-16}). In contrast, in the analyses of κ_2.0 data series, the success rates from

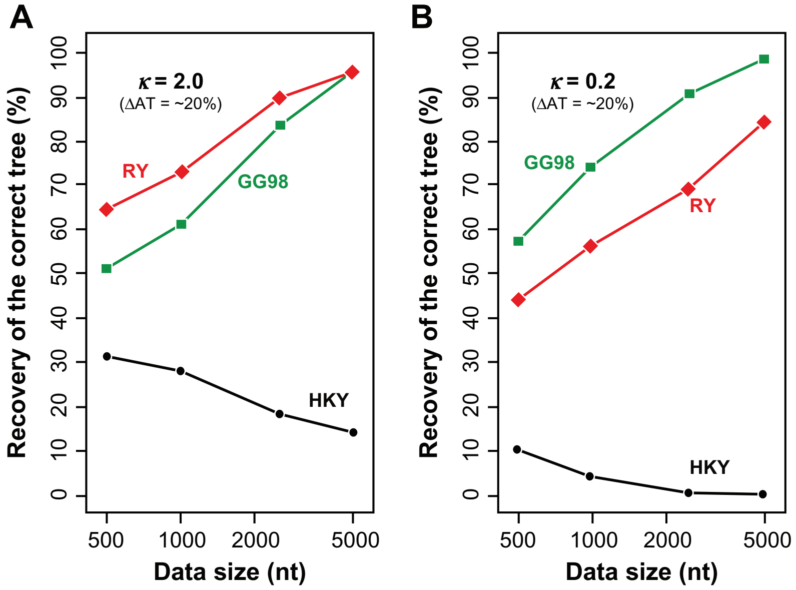


Figure 4. Impact of data size on the recovery rate of the correct tree. (**A**) We simulated four sets of sequence data with different sizes (500, 1000, 2500, and 5000 nt-long) with AT content across a tree of ~20% and transition/transversion ratio (κ) of 2.0. Five hundred replicates were simulated for each data point. The simulated data were subjected to the maximum-likelihood analyses with the HKY + Γ model (HKY; black circles) and the GG98 + Γ model (GG98; green squares). We also recoded the simulated data (comprising four nt characters, A, C, G, and T) into binary characters, purine (R) and pyrimidine (Y), and then subjected to the ML analysis with the CF + Γ model (RY; red diamonds). (**B**) The details are same as described in (A), but the sequence data were simulated with κ = 0.2.

RY-coding analyses were almost equal or greater than those from GG98 analyses (Fig. 4A). These results clearly suggest that the performance of RY-coding analysis can be altered by the evolutionary process that generated the sequence data of interest (eg, Ts/Tv ratio in this study).

Analysis of simulation data with complex base composition

We simulated an additional set of 4-taxon data with κ = 2.0 (1000 nt-long; 500 replicates). Unlike other simulated data analyzed in this study, neither frequencies of A and T nor those of C and G were set equal in these data. Slowly evolving Taxa 1 and 2 possess equal frequencies of the four bases, while rapidly evolving Taxa 3 and 4 possess approximately 45%, 25%, 13%, and 17% of A, T, G, and C, respectively (ΔAT = ~20%).

In this set of simulated data, purine (A and G) and pyrimidine (T and C) are equally contained in Taxa 1 and 2, while the ratio of purine to pyrimidine becomes almost 6:4 in Taxa 3 and 4. Thus, this compositional heterogeneity can introduce model misspecification to RY-coding analysis based on the CF + Γ model assuming the stationarity of R/Y composition across a tree. Similarly, the complex base composition simulated in sequence data cannot be modeled by the GG98 model, a non-homogeneous version of the T92 model,[31] in which assumes the frequencies of A and T, and those of C and G being equal. Indeed, the accuracies of RY-coding and GG98 analyses on this set of simulation data were lowered, dominantly recovering LBA tree (Fig. 5).

In theory, non-homogeneous models with more flexible assumption on base composition than GG98 model can improve the accuracy of the ML analysis. We then subjected the simulation data to the ML analysis with the nhHKY + Γ model, which allows the compositions of four bases to be independent. As anticipated, the accuracy of the ML analysis was greatly improved by applying the nhHKY + Γ model (Fig. 5).

Analysis of a real-world sequence data set

Prior to this study, only a single study has applied the GG98 model to the ML analysis of real-world data.[30] Unfortunately, due to the experimental setting of Herbeck et al[30] it was somewhat ambiguous whether the GG98 model suppressed the phylogenetic artifact from compositional heterogeneity.

We here subjected a real-world data set comprising five plastid-encoded genes, of which AT% varied

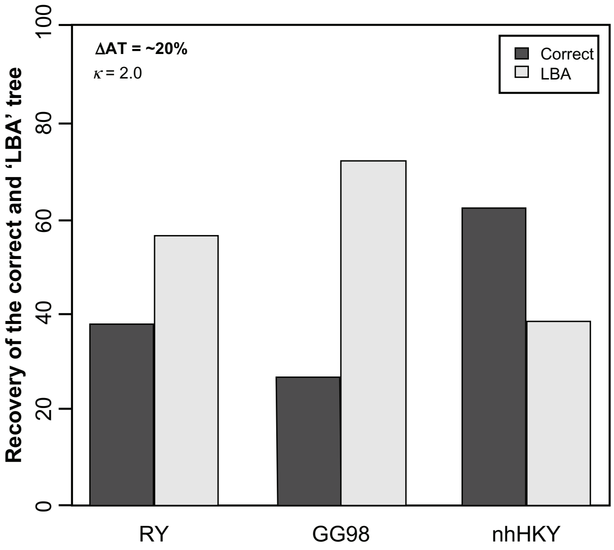


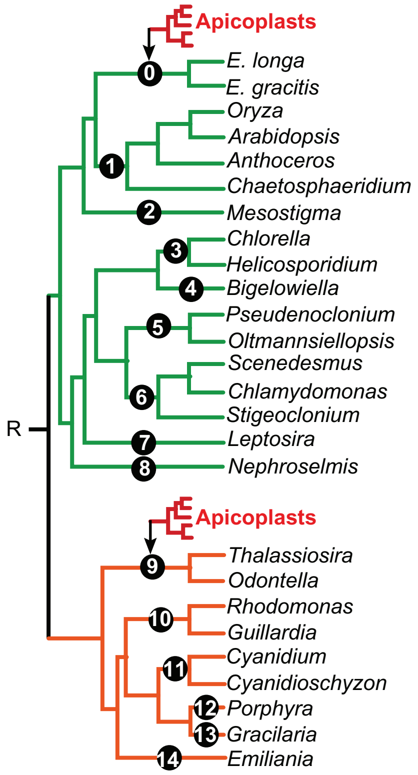
Figure 5. Impact of complex base composition on the maximum-likelihood analyses with RY-coding and non-homogeneous models. We simulated 1000 nt sequence data with AT content across a tree of ~20% and transition/transversion ratio (κ) of 2.0. Five hundred replicates were generated. The frequencies for A, C, G, and T were set equal in Taxa 1 and 2, while unequal base composition was applied to Taxa 3 and 4 (A = ~45%, T = ~25%, G = ~13%, C = ~17%). This set of simulated data was subjected to three different maximum-likelihood (ML) analyses—(i) 'RY-coding,' the ML analysis of the recoded data with the CF + Γ model; (ii) 'GG98,' the ML analysis with the GG98 + Γ model; (iii) 'nhHKY,' the ML analysis of the nhHKY + Γ model.
Note: The recovery of the correct and 'LBA' tree (see Fig. 1B) are shown as closed and open bars, respectively.

from 59.6% to 84.6% amongst the taxa considered, to the ML analysis with the GG98 + Γ model. The ML analysis of the 5-gene alignment with the homogeneous (HKY + Γ) model, which cannot take into account the ΔAT% across a tree, placed the apicoplast clade within green algal/green alga-derived plastids (Fig. 2A), and this tree topology is highly likely a phylogenetic artifact stemming from ΔAT%. Contrary to Herbeck et al,[30] we can directly assess whether the GG98 model suppresses the artifact from ΔAT% in the 5-gene alignment by assessing the position of the apicoplast clade. If the substitution process in the data was appropriately modeled, a tree representing the red origin of apicoplasts should be preferred over those representing the alternative hypothesis.

We examined the two competing hypotheses for the origin of apicoplasts by comparing the ML tree inferred from the HKY model-based analysis and 14 alternative trees, which are identical to the ML tree except for the position of the apicoplast clade (Fig. 6). In the tree comparison based on the lnL calculated with the HKY+ Γ model, the ML tree received the highest lnL score among the trees subjected to this comparison (Tree 0; Fig. 6), preferring the artifactual green origin of apicoplasts. In contrast, the GG98 + Γ model-based ML analysis supported the red origin of the apicoplasts—Trees 9–11 and 14, in which the apicoplast clade grouped with red alga/red alga-derived plastids, received higher lnL scores than any other trees, which represent the green origin of apicoplasts (Fig. 6). These results indicate that the GG98 + Γ model-based ML analysis successfully avoided a phylogenetic artifact stemming from ΔAT% in the 5-gene alignment.

Discussion

Compositional heterogeneity has been widely observed amongst molecular sequence data, and recognized as one of the important aspects in molecular sequence evolution for inferring accurate phylogenetic relationships.[4,9] To avoid or mitigate the artifacts stemming from compositional heterogeneity, there are two major choices available—cancelling compositional heterogeneity by character recoding, and accounting for compositional heterogeneity by applying a non-homogeneous model. Currently, non-homogeneous models are about to grow in popularity, mainly due to the limited number of phylogenetic programs that implement these complex models. On the other hand, character recoding seems a standard procedure for analyzing data with compositional heterogeneity—RY-coding for nt sequence data and 'Dayhoff-coding,' which converts 20 amino acid characters to four or six Dayhoff classes.[32] Nevertheless, the validities and limits of RY-coding and non-homogeneous models have not been fully examined in simulation studies. In the present study, we simulated the data series bearing 11 different degrees of compositional heterogeneity, and subsequently subjected these simulated data to RY-coding and GG98 analyses. We also examined whether the non-homogeneous (GG98) model overcame a putative artifact from heterogeneity of AT% in a real-world data set. Overall, both RY-coding and GG98 analyses showed superior performances than the control analyses with the homogeneous (HKY) model: Compositional heterogeneity in the simulated nt data had little impact on the success rate of RY-coding or GG98 analysis (Fig. 3A and B). We also observed that the GG98 + Γ model suppressed the phylogenetic artifact from the heterogeneity of AT% in the real-world sequences (Fig. 6).



	Apicoplast origin	HKY	GG98
Tree 0	Green	BEST (−51941.54)	5.65
Tree 1	Green	13.07	9.42
Tree 2	Green	18.11	9.98
Tree 3	Green	30.71	27.39
Tree 4	Green	26.45	24.16
Tree 5	Green	35.98	26.86
Tree 6	Green	35.98	28.27
Tree 7	Green	30.50	21.78
Tree 8	Green	26.68	16.66
Tree 9	Red	7.47	BEST (−50667.26)
Tree 10	Red	13.92	1.44
Tree 11	Red	14.13	2.03
Tree 12	Red	22.40	10.96
Tree 13	Red	21.05	9.91
Tree 14	Red	11.86	0.70

Figure 6. Tree log-likelihood comparison. The phylogram (left) is created from the tree topology shown in Figure 2A by pruning the entire apicoplast clade. The apicoplast clade was then regrafted to positions labeled 0–14 to generate the trees assessed in this comparison. For instance, Trees 0 and 9 were generated by regrafting the apicoplast clade to the branch leading to the clade of the green alga-derived plastids in two euglenids and that leading to the clade of the red alga-derived plastids in two diatoms, respectively. The root for the log-likelihood (lnL) calculation based on the non-homogeneous model is shown as "R." The table on the right provides the lnL value of the best tree among the 15 test trees, and the differences in lnL between the best tree and each of other trees. As shown in the second column (labeled as 'Apicoplast origin'), Trees 0–8 and 9–14 represent the 'green origin' and 'red origin' of apicoplasts, respectively. The values calculated with the HKY + Γ (homogeneous) model are listed in the third column (labeled as HKY), while those calculated with the GG98 + Γ (non-homogeneous) model are listed in the fourth column (labeled as GG98).

Nevertheless, it is noteworthy to mention that the performances of RY-coding analysis relative to that of GG98 analysis was largely altered by κ parameter setting in data simulation (Fig. 4A and B). We noticed that the overall site pattern was markedly different between the recoded κ_2.0 and κ_0.2 data (Fig. 7A). It is also noteworthy that the estimated branch lengths, particularly those for Taxa 3 and 4, calculated from the recoded κ_0.2 data were much longer than the corresponding values calculated from the recoded κ_2.0 data (Fig. 7B). Thus, the two differences shown in Figure 7A and B likely affected the relative performance of RY-coding analysis.

The results presented in this study clearly reinforce the importance of explicit incorporation of compositional heterogeneity in tree reconstruction. The maximum ΔAT% in the simulated and real-world (ie, 5-gene alignment) data examined here were ~20% and 25%, respectively, albeit some real-world data bear a higher magnitude of compositional heterogeneity (eg, ~37%[4] and ~50%[7]). Thus, severer LBA artifacts than what we observed here may be prevalent in homogeneous model-based analyses of real-world data. If compositional heterogeneity exists in the data of interest, we strongly recommend running both RY-coding and non-homogeneous model-based analyses. Despite the simplicity, RY-coding clearly improved the accuracy of tree reconstruction—the analyses after recoding can greatly mitigate the artifactual impact of compositional heterogeneity. However, we should be aware of the potential difficulties in this procedure: (i) the true phylogenetic signal in the original sequences can be erased by recoding, and (ii) the accuracy of RY-coding analysis, at least to some extent, depends on the substitution process that generated the data of interest (eg, Ts/Tv ratio; see Fig. 4A and B). We believe that non-homogeneous models are indispensable

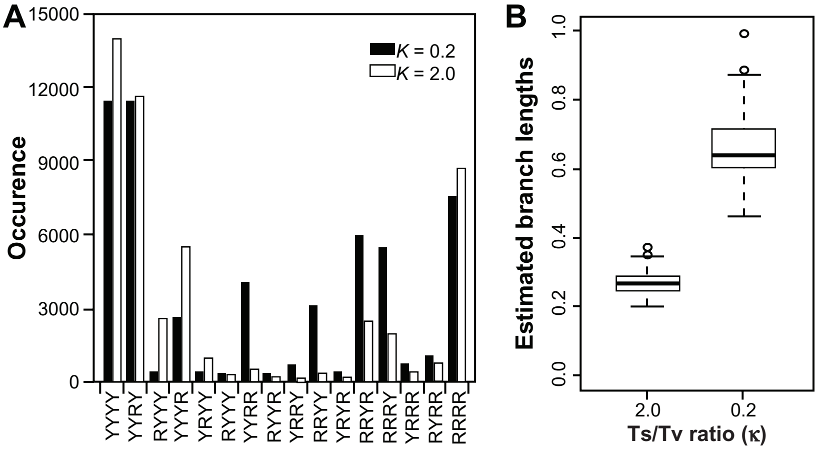


Figure 7. Impact of transition/transversion ratio in the data simulation on the maximum-likelihood analyses of the recoded data. (**A**) Difference in site pattern between the sequence data simulated with a transition/transversion ratio (κ) of 2.0 and those simulated with κ = 0.2 (shown as open and closed bars, respectively). We simulated a 50,000 nt-long simulated data, recoded into binary characters, purine (R) and pyrimidine (Y), and extracted the site pattern. (**B**) Lengths of the terminal branches leading to Taxa 3 and 4 estimated from the recoded data simulated with κ = 2.0 (left) and 0.2 (right). One thousand nt-long sequence data (500 replicates) were simulated and recoded into R and Y. We optimized the branch lengths of the correct tree, in which rapidly evolving Taxa 3 and 4 are separated (see Fig. 1B). Note that no 'correct' branch length is available for the results from RY-coding analysis, as the sequence data were not simulated as binary characters.

for analyzing molecular data bearing severe compositional heterogeneity since the non-homogeneous model-based analyses are supposed to be free from the potential issues in RY-coding analysis mentioned above.

Owing to the simple setting for data simulation, the GG98 model, which solely allows varying the AT content across a tree, was mostly used in the present study. However, the GG98 model cannot adequately account for a complex pattern of compositional heterogeneity across a tree in real-world data, in which the frequencies of A and T (or C and G) are unnecessarily equal. Our experiment evidently demonstrated that the violation of the assumption on base composition introduced LBA artifacts to the ML analysis with the GG98 model (Fig. 5). In practical, more complex and flexible non-homogeneous models than the GG98 model (eg, nhHKY model implemented in BppML[33]) may be useful for empirical phylogenetic analyses. At the same time, we need to judge what kind of non-homogeneous model is most appropriate for describing the substitution process in the data of interest to avoid over-fitting. Nevertheless, to our knowledge, no program for model selection can take non-homogeneous models into account. Thus, more advanced programs for model selection are also indispensable before starting applying non-homogeneous models to phylogenetic analyses of various empirical data.

Author Contributions

Conceived and designed the experiments: SAI, YI, TH. Analysed the data: SAI. Wrote the first draft of the manuscript: SAI. Contributed to the writing of the manuscript: SAI. Agree with manuscript results and conclusions: SAI, YI, TH. Jointly developed the structure and arguments for the paper: SAI, YI, TH. Made critical revisions and approved final version: SAI, YI, TH. All authors reviewed and approved of the final manuscript SAI, YI, TH.

Funding

This work was supported by grants from the Japanese Society for the Promotion of Science (Nos. 21370031 and 22657025 awarded to YI; 2347038 and 23405013 awarded to TH), and those from the Ministry of Education, Culture, Sports, Science and Technology (Nos. 23117006 and 23117005 awarded to YI and TH, respectively).

Acknowledgements

We thank Drs. Ryoma Kamikawa and Akifumi Tanabe (University of Tsukuba, Japan) for technical advice and critical comments on the manuscript.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Philippe H. Opinion: long branch attraction and protest phylogeny. *Protist*. 2000;151:307–16.
2. Huelsenbeck JP. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol Biol Evol*. 1995;12:843–9.
3. Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Roger JS. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol*. 2001;50:525–39.
4. Hasegawa M, Hashimoto T. Ribosomal RNA trees misleading? *Nature*. 1993;361:23.
5. Foster PG, Hickey DA. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol*. 1999;48:284–90.
6. Chang BSW, Campbell DL. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Mol Biol Evol*. 2000;17:1220–31.
7. Mooers A, Holmes EC. The evolution of base composition and phylogenetic inference. *Trends Ecol Evol*. 2000;15:365–9.
8. Tarrio R, Rodríguez-Trelles F, Ayala FJ. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. *Mol Biol Evol*. 2001;18:1464–73.
9. Galtier N, Gouy M. Inferring phylogenies from DNA sequences of unequal base compositions. *Proc Natl Acad Sci U S A*. 1995;92:11317–21.
10. Conant GC, Lewis PO. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. *Mol Biol Evol*. 2001;18:1024–33.
11. Ho SYW, Jermiin LS. Tracing the decay of the historical signal in biological sequence data. *Syst Biol*. 2004;53:623–37.
12. Jermiin LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol*. 2004;53:637–43.
13. Lockhart PJ, Steel MA, Hendy MD, Penny D. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*. 1994;11:605–12.
14. Steel MA. Recovering a tree from the leaf colourations it generates under a Markov model. *Appl Math Lett*. 1994;7:19–23.
15. Rodriguez-Trelles F, Tarrio R, Ayala FJ. Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *J Mol Evol*. 2000;50:1–10.
16. Singh ND, Arndt PF, Clark AG, Aquadro CF. Strong evidence for lineage and sequence specificity of substitution rates and patterns in *Drosophila*. *Mol Biol Evol*. 2009;26:1591–605.
17. Inagaki Y, Simpson AGB, Dacks JB, Roger AJ. Phylogenetic artifacts can be caused by leucine, serine, and arginine codon usage heterogeneity: dinoflagellate plastid origin as a case study. *Syst Biol*. 2004;53:582–93.
18. Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol*. 2007;56:389–99.
19. Phillips MJ, Penny D. The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol Phylogenet Evol*. 2003;28:171–85.
20. Phillips MJ, Delsuc F, Penny D. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*. 2004;21:1455–8.
21. Cavender JA, Felsenstein J. Invariants of phylogenies in a simple case with discrete states. *J Classif*. 1987;4:57–71.
22. Galtier N, Gouy M. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol*. 1998;15:871–9.
23. Strope CL, Abel K, Scott SD, Moriyama EN. Biological sequence simulation for testing complex evolutionary hypotheses: indel-Seq-Gen Version 2.0. *Mol Biol Evol*. 2009;26:2581–93.
24. Hasegawa M, Kishino H, Yano T. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 1985;22:160–74.
25. Swofford DL. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4 beta 10. Sinauer Associates, Sunderland, Mass; 2003.
26. Galtier N, Tourasse N, Gouy M. A nonhyperthermophilic common ancestor to extant life forms. *Science*. 1999;83:220–1.
27. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 2010;59(3):307–21.
28. Janouškovec J, Horák A, Oborník M, Lukeš J, Keeling PJ. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad Sci U S A*. 2010;107:10949–54.
29. Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H. Additional molecular support for the new chordate phylogeny. *Genesis*. 2008;46:592–604.
30. Herbeck JT, Degnan PH, Wernegreen JJ. Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (γ-Proteobacteria). *Mol Biol Evol*. 2004;22: 520–32.
31. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*. 1992;9:678–87.
32. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci U S A*. 2008;105:20356–61.
33. Dutheil J, Boussau B. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. *BMC Evol Biol*. 2008;8:255–5.

Supplementary Figures

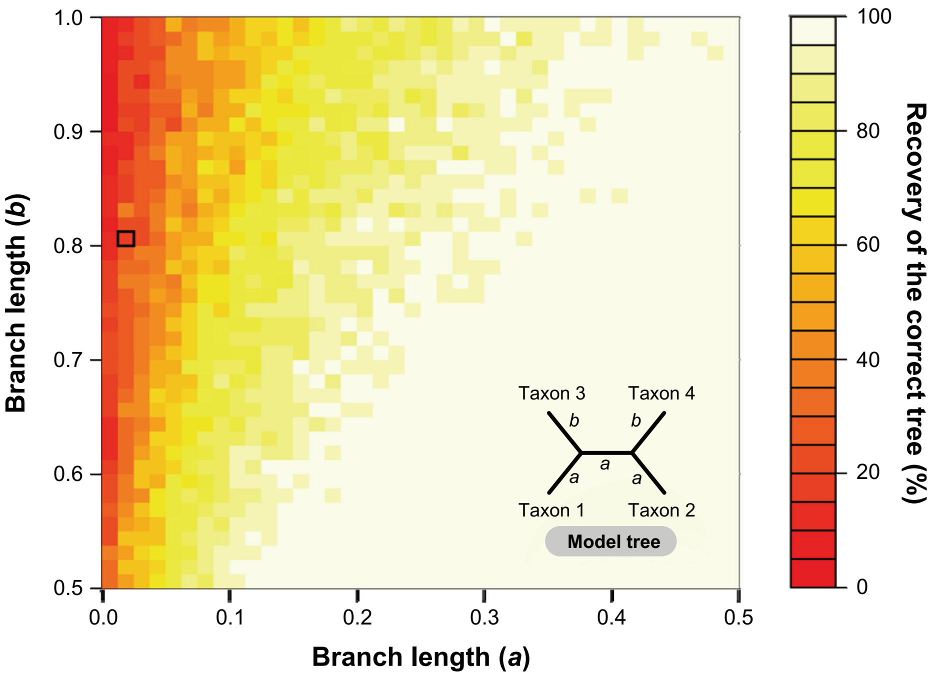


Figure S1. Impact of the branch lengths on the recovery ratio of the correct tree in the maximum-likelihood analysis with HKY + Γ model. We simulated 1,000 nucleotide-long sequence data with AT content across tree of ~20% and transition/transversion ratio (κ) of 2.0 based on the 4-taxon model tree. 40 * 40 combinations of branch lengths of *a* and *b* of the model tree were examined. For each combination, we analyzed 100 replicates. The color of each cell in the matrix indicates the success rate.

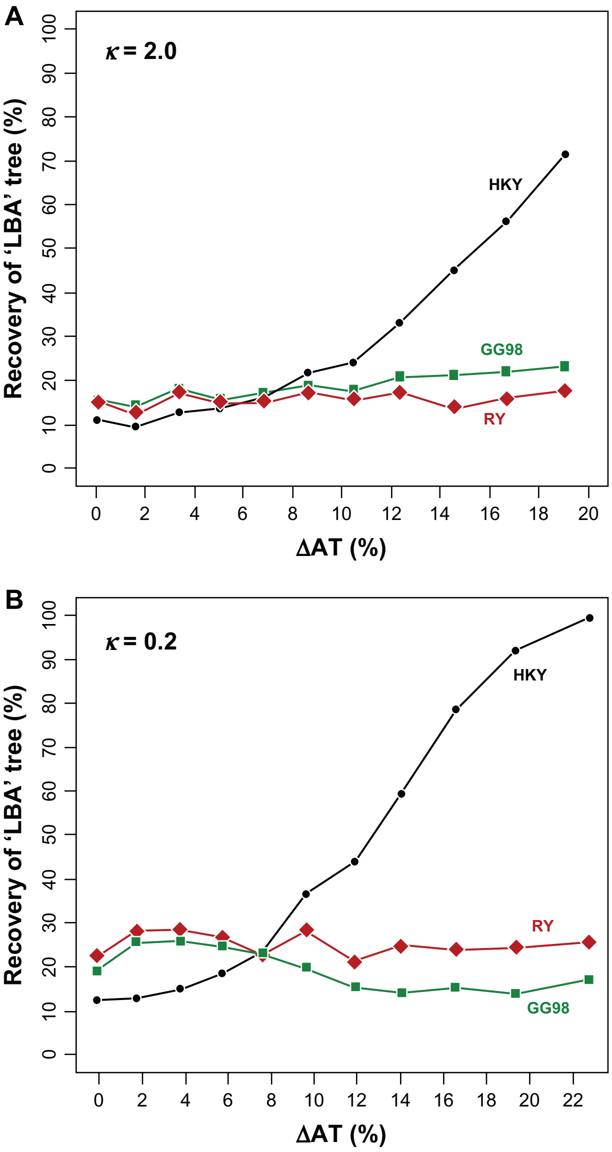


Figure S2. Impact of the difference in AT content across a tree on the recovery rate of 'LBA' tree, in which rapidly-evolving Taxa 3 and 4 group together (see Fig. 1B).
Note: The details of these figures are same as those in Fig. 3A and B, except we plotted the recovery rates of LBA tree here.