How to Reconstruct Accurate Phylogenetic Trees from Nucleotide Sequence Data with Extraordinary Compositional Bias: Assessment of the Performance of Data-Recoding Methods and Non-Homogeneous Models

A Dissertation Submitted to
the Graduate School of Life and Environmental Sciences,
the University of Tsukuba
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Sciences
(Doctoral Program in Biological Sciences)

Sohta ISHIKAWA

TABLE of CONTENTS

ABSTRACT

Phylogenetic analyses have been widely used to infer evolutionary relationships amongst life forms, based on molecular data such like nucleotide and amino-acid sequences. In phylogenetic analyses, 'homogeneous' substitution models, which assume the homogeneity of base or amino-acid composition across lineages, are generally applied. However, the assumption of homogeneous models is often violated by the heterogeneity of base or amino-acid composition in real-world sequences. Especially, the heterogeneity of adenine plus thymine (AT) content in nucleotide sequences is widely recognized to interrupt accurate inference of evolutionary history in the analyses with homogeneous models. To avoid or mitigate phylogenetic artifacts stemming from the heterogeneity of AT content, I here focused on the two approaches, 'data-recoding' methods and 'non-homogeneous (NH)' substitution models.

In chapter 1 and chapter 2, I demonstrated a comprehensive study to assess the robustness of a data-recoding method, 'RY-coding,' and NH models, by analyzing simulated and real-world sequence datasets with various degrees of the heterogeneity of AT content. From my results, RY-coding and NH models successfully improved phylogenetic inferences compared to homogeneous models, even under the presence of ~20% of AT content heterogeneity in both simulated and real-world sequence datasets. Nevertheless, I revealed that the accuracy of RY-coding-based analysis can be affected by i) the substitution process that generated the sequence data, ii) the level of the heterogeneity of base composition, and iii) the loss of true phylogenetic signal due to recoding procedure. On the other hands, NH models were revealed to be free from such difficulty of the data-recoding method.

Phylogenetic inferences with NH models, however, can be computationally intense because an enormous amount of model parameters need to be optimized. In chapter 3, I performed a methodological approach to reduce the computational time for the phylogenetic analyses with NH models. I applied two parallel computing methods, OpenMP and MPI, to a phylogenetic program for the maximum-likelihood inference with a NH model. The parallelized program achieved suitable speeding-up up to 64 computational nodes and 1,024 CPU cores on a supercomputer system, 'T2K-Tsukuba.'

In conclusion, I discuss the pros and cons of the data-recoding method and NH models based on the results obtained here. The goal of the present study is to provide a guideline to properly use these two methodologies in future phylogenetic analyses, with diverse empirical sequence datasets bearing a variety of compositional heterogeneity.

ABBREVIATIONS

AT content	Adenine plus Thymine content
LBA	Long Branch Attraction
lnL	log-likelihood
ML	Maximum-Likelihood
NH	Non-Homogeneous
Ts/Tv	Transition/Transversion

I. GENERAL INTRODUCTION

I-1 Phylogenetic analyses

I-1-1 Phylogenetic trees

One of the most important purposes in evolutionary biology is to investigate the relationship amongst life forms and elucidate the evolutionary process from the common ancestor to existent organisms. Phylogenetic analyses (phylogenetic inferences), which infer the phylogenetic relationships among a group of organisms based on molecular data such as nucleotide and protein sequences, have played a key role to accomplish above goals in evolutionary studies. In phylogenetic analyses, any taxonomic categories such as species, order, family, etc., or genes in several cases, can be dealt as Operational Taxonomic Unit (OTU) or taxon (taxa, plural form). Of note, I here use taxon and taxa. The relationships between taxa are illustrated by means of a phylogenetic tree, which is composed of internal/terminal nodes and branches. The terminal nodes in a phylogenetic tree indicate extant taxa and internal nodes represent ancestral taxa. The branch connects two adjacent nodes, defining the ancestor-descendant relationships. Thus, the evolutionary relationships among taxa can be described as a tree-like pattern (topology). A tree can be completely bifurcated if all nodes have only two immediate descendant lineages, but multifurcated if a node has more than two immediate descendant lineages. Phylogenetic trees can be either rooted or unrooted. The rooted tree has a root, which means the common ancestor of all taxa under study, and a unique path leads from the root to any other nodes. The direction of each path corresponds to the evolutionary time. In contrast, the unrooted tree only specifies the relationships among taxa with no time direction.

I-1-2 Methods for inferring phylogenetic trees

A phylogenetic tree can be inferred from molecular sequences retrieved from extant taxa. They can be aligned into the single alignment in which we can find the difference of characters (i.e., bases or amino-acids) among taxa on each position. Such differences are caused by substitutions that occurred on each sequence during its independent evolutionary process. Hence, the tree topology and corresponding branch lengths (i.e., the average number of substitutions on each position of a sequence), can be inferred by analyzing the substitution processes that generated the observed sequence alignment.

There are several methods to infer the phylogenetic trees from molecular sequence data. The maximum-parsimony (MP) method infers the minimum number of substitutions that are required to explain all observed differences among extant sequences, considering a particular tree topology [1, 2]. The MP method counts the number of substitutions for each of the possible tree topology and selects the one showing the smallest number of substitutions as the most optimal tree. The distance-matrix (DM) method calculates a genetic distance, which is defined as a number of substitutions per position per unit time, between all pairs of extant sequences. Then, the phylogenetic tree is reconstructed based on the matrix of distances, using various algorithms for clustering taxa, including the un-weighted pair-group method with arithmetic mean (UPGMA) [3] and the neighbor-joining (NJ) method [4].

In this thesis, I focus particularly on the maximum-likelihood (ML) method [5], which is one of the most popular approaches to infer phylogenetic trees. In the phylogenetic analyses based on the ML method (henceforth designated as the ML

analyses), each extant sequence is assumed to have evolved following a stochastic process for substitutions, called as 'Markov process.' Based on a Markov process, the substitution process for molecular sequences can be described by a rate matrix called as 'substitution model,' where the rates for character *i* (e.g., four bases or 20 amino-acids) being replaced by character *j* in an instantaneous time are defined by model parameters. Using substitution models, we can calculate the probability of substitutions that occurred in a particular evolutionary time (*t*) on a branch of a tree. Then, a likelihood of a given tree topology can be obtained by multiplying whole substitution probabilities among branches. Note that the likelihood is a counterpart of the probability of observed sequence data with respect to branch lengths and model parameters.

For instance, suppose that an unrooted tree as shown in Fig. 1, which is composed of four terminal branches (t_1, t_2, t_3, and t_4) and one internal branch (t_5). We here observe bases (*p*, *q*, *r*, *s*) on the position *h* of the nucleotide sequence data of under study (*D*). Each observed base is plotted on each terminal node in Fig. 1. In most cases, the bases *i* and *j* on internal nodes (ancestral sequences) are not known. Therefore, the whole possibilities for the bases in internal nodes are considered. The likelihood of the tree in Fig. 1 on the position *h* can be calculated as described below.

$$L(\theta|D_h) = \sum_{i=A,T,G,C} \{ \quad \pi_i p_{ip}(t_1) p_{iq}(t_2) \sum_{j=A,T,G,C} p_{ij}(t_5) p_{jr}(t_3) p_{js}(t_4) \quad \}$$

The parameter π_i represents the frequency of the base *i* on the corresponding internal node (Fig. 1). π_i can be estimated from the entire sequence alignment. The $p_{ij}(t)$ denotes the probability for the substitution from *i* to *j* in the evolutionary time defined by the length of the corresponding branch *t*. $p_{ij}(t)$ can be calculated based on a given nucleotide substitution model. Since each position is assumed to evolve independently,

the likelihood of the tree for whole sequence data, $L(\Theta|D)$, can be obtained by multiplying the likelihoods calculated position by position. The ML method estimates the model parameters and branch lengths which maximize the likelihood of a given tree. In the same way, we can also calculate the likelihoods of alternative tree topologies. Finally, the ML method selects the topology showing the highest likelihood value as the most optimal tree (the ML tree). Note that we generally report the log-likelihoods (ln*L*s), as likelihoods are extremely small.

I-2 Artifacts of Phylogenetic inference

I-2-1 Long-Branch Attraction

In phylogenetic analyses, two distantly related, but rapidly evolving (long-branch) taxa often erroneously group together owing to long-branch attraction (LBA) [6]. Such phylogenetic artifacts caused by LBA have been recognized as one of the major sources that mislead the accurate inference of the evolutionary relationships among diverse organisms [7–9]. Pioneering studies based on simulated data have shown that the susceptibility to LBA artifacts differs amongst tree reconstruction methods—the MP and DM methods are sensitive to, but the ML method is in theory robust against LBA artifacts [10, 11]. This ideal property of the ML method, however, collapses under conditions such as ‘model misspecification,’ where the substitution model does not appropriately describe the substitution process that generated the sequence data of interest. As the precise substitution 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 [12, 13].

I-2-2 Compositional heterogeneity

The compositional bias in sequence data, i.e., the heterogeneity of base or amino-acid composition among sequences, has been regarded as one of the most important sources for model misspecification [14, 15]. I here bring up the compositional bias in nucleotide (nt) sequences, which are the most fundamental materials for phylogenetic inferences. As base composition varies amongst genes, or even genomes, the heterogeneity of base composition is likely ubiquitous in nt alignments [16]. On the other hand, widely used nt substitution models, which are based on the stationary Markov process across tree, assume the homogeneity of base composition among sequences; that is, all sequences are supposed to have evolved following same base frequencies, which are estimated from the entire alignment [17]. Such assumption in 'homogeneous' substitution models, however, can be violated by the 'non-homogeneous' sequence evolution where each sequence has evolved following independent base frequencies. Therefore, analyzing nt data bearing compositional bias under homogeneous model conditions introduces significant model misspecification to tree reconstruction, resulting in severe phylogenetic artifacts [18].

In particular, adenine + thymine (AT) content (or guanine + cytosine (GC) content) have been reported to vary at genome level within or across groups of organisms. For instance, prokaryotes are known to have the widest diversity of genomic AT contents from 23% to 83.5% [19–21]. The range in genomic AT content in eukaryotes is also variable from AT-rich (AT > 50%) to AT-poor (AT < 50%) [22–27]. Plastid and mitochondrial genomes sequenced to date also show relatively high, but

various degrees of AT content [28–31]. Thus, the heterogeneity of AT content can be considered as a major source of compositional bias in the analyses of nt sequence data. Importantly, it was revealed that biased AT contents are strongly related with rapid evolutionary rates [32], implying that LBA artifacts in phylogenetic analyses can be enhanced by AT content bias. Indeed, analyses of both real-world and simulated sequence data have shown that the ML analyses based on homogeneous models, which assume the homogeneity of AT content across taxa, can misleadingly grouped unrelated taxa bearing rapid evolutionary rates and similar AT contents [33–37].

I-3 Data-recoding method and non-homogeneous substitution models

Heterogeneity of AT content has been widely observed in empirical sequence data for phylogenetic analyses, and recognized as one of the most important aspects in molecular sequence evolution for inferring accurate phylogenetic relationships [38–41]. To avoid or mitigate phylogenetic artifacts stemming from AT content heterogeneity across tree, there are two major choices available—cancelling compositional bias by a data-recoding method, and accounting for compositional heterogeneity by applying the non-homogeneous (NH) substitution model. In the former method, the variation of AT (or GC) content in nt alignments can be efficiently homogenized by recoding four characters, A, C, G, and T, into purine (R; A or G) or pyrimidine (Y; C or T). This ‘RY-coding,’ which can be coupled with an ML method using a substitution model for two-state characters [42], was initially proposed to prevent the putative artifact in the analyses of mammalian nt sequence datasets [43, 44]. The latter method, NH models, can theoretically relax the assumption of the homogeneity of base composition by allowing the model parameters to vary in branch-by-branch fashion across a tree [45–

50]. Therefore, the non-homogeneous sequence evolution can be appropriately described under NH model conditions. Especially, a NH model provided by Galtier and Gouy (1998), henceforth designated as ‘GG98 model,’ can explicitly take the heterogeneity of AT content into account by implementing free parameters for estimating equilibrium AT content on each branch of a tree. Thus, each nt sequence is assumed to have evolved under different degrees of AT content at each point of its evolutionary path from the common ancestor [49].

I-4 Purpose of this study

RY-coding and GG98 model have been applied in several pioneer studies tackling the accurate phylogenetic inference from real-world sequence datasets which exhibit severe AT content heterogeneity [41, 43, 44, 51–55]. In these data analyses, both two methods successfully suppressed phylogenetic artifacts, compared to the analyses with homogeneous models. Nevertheless, there is still room for argument on basic properties of RY-coding and GG98 model. I here propose two central questions, i) how efficiently these two methods can recover the true phylogenetic relationship, and ii) to what extent of the AT content heterogeneity they can tolerate.

To discuss the above issues, I here conducted a comprehensive analysis to assess the performance of the ML analyses incorporating RY-coding and GG98 model, based on simulated and real-world sequence datasets. I also conducted a computational effort to reduce the computational cost of GG98 model, which potentially limits the application of the model to the analyses of large-scale sequence datasets. Summarizing the results obtained here, I finally discuss the usage of the data-recoding method and NH models in depth, aiming to apply them into future phylogenetic analyses for wide

range of empirical sequence datasets bearing significant compositional heterogeneity.

II. PERFORMANCE EVALUATION FOR RY-CODING AND GG98 MODEL WITH SIMULATED DATASETS

II-1 Introduction

RY-coding coupled with an ML method can mitigate the heterogeneity of AT content by character recoding, and is believed to ameliorate the accuracy of phylogenetic inferences [43, 44]. Nevertheless, there are some potentially unclarified issues in this procedure. First, it 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 a homogeneous substitution model for two-state character proposed by Cavender and Felsenstein [42], it is naïve to assume that the ML inferences from the recoded alignments are completely liberated from the phylogenetic artifacts from compositional heterogeneity. Second, the recoding procedure may discard informative transition substitutions (A↔G or T↔C) in the original alignments, which may reduce the resolution of the true phylogenetic relationship. Importantly, the efficacy of RY-coding, as well as its potential limitation, remains uncertain because no simulation study exhaustively assessing the above concerns is available so far.

On the other hand, GG98 model can explicitly take the heterogeneity of AT content across a tree into account by allocating its model parameters in branch-by-branch fashion [49]. A study based on simulated nt data with biased base composition evidently showed that the accuracy of a distance matrix (DM) based method was greatly improved by GG98 model [50]. Analysis with GG98 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 GG98 model is anticipated to be much more robust against typical LBA artifacts than any DM-based methods. However, to date, the robustness of ML inferences under GG98 model conditions has not yet been examined in detail by analyzing simulated data.

I here present the results from the de facto first simulation study assessing the performance of an ML method incorporating RY-coding and that with GG98 model. Simulated nt sequence datasets bearing various degrees of the heterogeneity of AT content were subjected to the two types of ML analyses. My study clearly indicated that the ML analyses incorporating RY-coding and GG98 model (henceforth designated as RY-coding and GG98 analyses, respectively) were more robust against the LBA artifact stemming from AT content bias than the ML analysis with a homogeneous substitution model, which cannot take compositional heterogeneity into account. Nevertheless, my closed investigation revealed the potential pitfalls of both RY and GG98 analyses. The performance of RY analysis appeared to be largely affected by the substitution process used for sequence simulation. Likewise, the inference from GG98 analysis could be significantly misled when the complex pattern of compositional heterogeneity violated the assumption of the model.

II-2 Materials and Methods

II-2-1 Data Simulation

Nucleotide sequence data was generated by Monte Carlo simulation, using indel-Seq-Gen Version 2.0 [56], based on a 4-taxon model tree described below (Fig. 2A). I simulated 500 replicates for each data point. 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 0.8 (*a* and *b* in Fig. 2A). For data simulation, the ancestral sequences were randomly generated at the root (R in Fig. 2A), and each tip sequence was then simulated according to the given branch lengths. The substitution process was modeled with the HKY model [57], incorporating rate heterogeneity across sites approximated by a discrete gamma (Γ) distribution [58] with four categories (henceforth designated as HKY + Γ model). The κ parameter for Ts/Tv ratio [59] and the shape parameter α for a Γ distribution were set to 2.0 and 0.8, according to Galtier and Gouy (1995) [49]. I 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 performance 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 (i.e. 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. 2A). The above procedure enabled me 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. I analyzed the simulated datasets with 11 variations of the difference of AT% between slowly evolving Taxa 1 and 2, and rapidly evolving Taxa 3 and 4 (henceforth designated as ΔAT%). The frequencies of A and T and those of C and G were set equal unless I specifically mention. The settings for base frequency in the data simulation, and the average AT% achieved in the resultant simulated data are presented in Table 1.

II-2-2 Data Analyses

I ran three different ML analyses in the present study. First, the simulated data, comprising four bases, 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 = 0.2–2.0, and α = 0.8), but base frequencies were estimated from the entire data. I also analyzed the simulated data recoded by RY-coding [44, 60]. The recoded data (comprising two-state characters) were then analyzed with the model of Cavender and Felsenstein [42] 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 base (and R/Y) composition. I used PAUP* 4.0b [61], for the ML analyses with the two homogeneous models.

Finally, I subjected the simulated nt sequence data to the third ML analysis with GG98 model [49] incorporating rate heterogeneity across sites approximated by a discrete Γ distribution (GG98 + Γ model), which was implemented in NHML 3.0 [53]. In this non-homogeneous model, the parameters for Ts/Tv ratio and the Γ distribution were estimated from the entire data and fixed across a tree, but the parameter for AT content was allowed to vary in a branch-by-branch fashion. I exhaustively searched for the ML tree by *eval_nh* program packaged in NHML. 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 (henceforth designated as nhHKY + Γ model). I used BppML program implemented in Bio++ 0.8.0 [48] for the

data analyses with the nhHKY + Γ model.

II-3 Results

II-3-1 Impact of compositional heterogeneity—HKY analysis

The HKY model assumes the stationarity of the substitution process (i.e., homogeneous), and ΔAT% in the simulated data cannot be adequately accounted for. Henceforth here, I designate the HKY model-based ML analysis as 'HKY analysis.' On the basis of Jermiin et al. (2004) and Ho and Jermiin. (2004), I expected that 'LBA' tree (center in Fig. 2B), in which rapidly evolving Taxa 3 and 4 erroneously grouped together, was preferentially recovered in HKY analysis of the data bearing large ΔAT%.

First, as a preliminary analysis, 1,000 nt sequence datasets simulated under 1,600 combinations of branch lengths, with *a* (Fig. 2A) ranging from 0.0125 to 0.5, and *b* (Fig. 2A) ranging from 0.5 to 1.0, were analyzed. Fig. 3 shows the difference of recovery ratio of the correct tree in HKY analyses under the 1,600 combinations of branch lengths, with AT content across tree of ≈20% and Ts/Tv ratio (κ) of 2.0. I determined specific branch lengths *a* and *b* (a = 0.025, b = 0.8; see II-2-1), under which HKY analysis showed significantly low recovery ratio of the correct tree (left in Fig. 2B) due to the LBA attraction and the heterogeneity of AT content (boxed area in Fig. 3).

In the analysis of 1,000 nt-long data simulated with κ = 2.0 and fixed branch lengths of *a* and *b* (henceforth designated as 'κ_2.0 data'), the recovery rate of the correct tree gradually decreased along with the increment of ΔAT% (black circles in Fig. 4A). On the other hand, LBA tree was dominantly yielded in the analyses of the data with high ΔAT% (black circles in Fig. 5A). 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 $\kappa = 0.2$ (henceforth designated as 'κ_0.2 data;' black circles in Figs. 4B and 5B). 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.

I additionally tested how the performance of HKY analysis was affected by the Ts/Tv ratio in data simulation. Five sets of 1,000 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 Fig. 4C, the analysis of κ_2.0 data yielded the highest success rates (≈30%), while the correct tree was recovered at less than 10% in the analyses of the data simulated with $\kappa < 2.0$.

II-3-2 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 [44, 60, 62]. However, there is a (potentially large) room for argument on whether this procedure can truly help in inferring the correct tree. In this study, both κ_2.0 and κ_0.2 data series bearing ΔAT of 0–20% were subjected to the RY-coding analysis.

I firstly checked whether the recoding procedure erased the compositional heterogeneity simulated in κ_2.0 and κ_0.2 data. As shown in Table 2, regardless of the setting for AT% in Taxa 3 and 4 in original simulated data, as well as Ts/Tv ratio,, the difference of purine (R) between Taxa 1 and 2, and Taxa 3 and 4 (ΔR%) was fixed to about 2% in the recoded data. As almost no compositional heterogeneity existed in recoded data, 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 Figs. 4A and 4B). 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 (red diamonds in Figs. 5A and 5B). The success rate of RY-coding analysis remained higher irrespective of Ts/Tv ratio (56–70%; red diamonds in Fig. 4C), compared with that of HKY analysis (black circles in Fig. 4C). I successfully provide the first simulation results that indicate that RY-coding largely improved the phylogenetic inferences of sequence data with compositional heterogeneity.

II-3-3 Impact of compositional heterogeneity—GG98 analysis

The non-homogeneous GG98 model proposed by Galtier and Gouy (1998) [50] allows different AT% on different branches. 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 [35, 54, 63]. Nevertheless, although simulation study by Galtier and Gouy (1995) [49] showed that GG98 model drastically improved the accuracy of a DM-based analysis, the performance of GG98 model-based ML analysis (henceforth here designated 'GG98' analysis) has not been fully tested. In the present study, I examined how efficiently 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. 4A), while the recovery of LBA tree was suppressed (<23%; Fig. 5A). In the GG98 analysis of κ_0.2 data series, ΔAT% had little impact on the success rate (63–72%; green squares in Fig. 4B) and the false rate (14–26%; green squares in Fig. 5B). 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. 4C). These are the first simulation results indicating that the parallel shifts of AT content in nt sequence data could be robustly tolerated in NH model-based ML analysis.

II-3-4 Impact of data size

I simulated 500, 1,000, 2,500, and 5,000 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 Figs. 6A and 6B. Regardless of κ parameter, the success rate of HKY analysis appeared to be negatively correlated with data size (black circles in Figs. 6A and 6B). The analyses of the largest κ_2.0 and κ_0.2 data (i.e., 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 increasing data size.

In contrast, 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 Figs. 6A and 6B). 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 similarly improved by increasing data size (up to 95% and 98%, respectively; green squares in Figs. 6A and 6B). 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.

II-3-5 GG98 analysis versus RY-coding analysis

Both RY-coding and GG98 analyses were robust against ΔAT% in the simulated data (Figs. 4 and 5), and their success rates displayed positive correlation with data size (Fig. 6). 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. 6B). I statistically compared the success rates of 500 simulation trials from RY-coding and GG98 analyses for 500, 1,000, 2,500, and 5,000 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}). On the other hand, in the analyses of κ_2.0 data series, the success rates from RY-coding analyses were almost equal or greater than those from GG98 analyses (Fig. 6A). 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 (e.g., Ts/Tv ratio in this study).

II-3-6 Analyses with more complex pattern of compositional heterogeneity

I simulated an additional set of 4-taxon data with $\kappa = 2.0$ (1,000 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, which is a non-homogeneous version of the TN92 model [64] assuming 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 significantly lowered, dominantly recovering LBA tree (Fig. 7).

In theory, NH models with more flexible assumption on base composition than GG98 model can improve the accuracy of the ML analysis. Therefore, I subjected the simulation data to the ML analysis with the nhHKY + Γ model, which allows the frequencies of three of the four bases to be independent. As anticipated, the accuracy of the ML analysis was greatly improved by applying the nhHKY + Γ model (Fig. 7).

II-4 Discussion

The validities and limits of RY-coding and GG98 model have not been fully examined by simulation with a variety of experimental settings. In the present study, I simulated nt sequence data series bearing 11 different degrees of the heterogeneity of AT content across taxa, and subsequently subjected them to RY-coding-based and GG98 model-based analyses. Overall, both RY-coding and GG98 model analyses showed superior performances than the control analyses with a homogeneous (HKY) model. The maximum ΔAT% examined here were ≈20%, albeit some real-world data bear a higher magnitude of the heterogeneity of AT content across lineages (e.g., ~37% in [33] and ~50% in [16]). Thus, severer artifacts than what I observed here may be prevalent in

real-world data analyses based on homogeneous models. However, results from my simulation provide strong evidence to support that the degree of the heterogeneity of AT content had little impact on the success rate of RY-coding or GG98 analysis (Figs. 4A and 4B).

Nonetheless, 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 (Figs. 4 and 6). I noticed that the overall site pattern was markedly different between the recoded κ_2.0 and κ_0.2 data (Fig. 8A). 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. 8B). Thus, the two differences observed on the analyses of the recoded data series (Figs. 8A and 8B) likely led to the difference on the recovery rate of the correct tree (Figs. 4A and 4B). It is generally assumed that there is a universal bias in favor of transitions over transversions [59]. However, a previous work has revealed that such 'universal' rule cannot be applied to some real-world sequence [65], largely implying that the performance of RY-coding-based analysis can be affected by the Ts/Tv ratio and produce phylogenetic artifacts.

In contrast, GG98 model is perhaps more efficient than RY-coding method, since the GG98 model-based analyses are supposed to be free from the potential issues in the data-recoding procedure mentioned above. However, I should point out that GG98 model may not adequately account for complex patterns of compositional heterogeneity among real-world sequences, in which the frequencies of A and T (or C and G) are unlikely equal. My experiment evidently demonstrated that the violation of the assumption on base composition introduced phylogenetic artifacts to the ML

analysis even with the GG98 model (Fig. 7). In such case, more complex and flexible NH models than the GG98 model (e.g., nhHKY model implemented in BppML [48]) may be useful for empirical phylogenetic analyses).

Finally, the results presented in this simulation study clearly reinforce the importance of explicit incorporation of compositional heterogeneity in phylogenetic inferences.

III. PERFORMANCE EVALUATION FOR RY-CODING AND GG98 MODEL WITH A REAL-WORLD SEQUENCE DATASET

III-1 Introduction

In my simulation study in chapter 2, both RY-coding and GG98 model certainly showed their ability to reconstruct accurate phylogenetic trees under the presence of various degrees of AT content heterogeneity. Nevertheless, my closed investigation also revealed the potential pitfalls of RY-coding and GG98 model. The performance of RY-coding analysis appeared to be largely affected by substitution processes that generated the data of interest (Figs. 4A, 4B, and 7). Likewise, the phylogenetic inferences with GG98 model may be misled when the pattern of base composition in the data violated the assumption of the model (Fig. 7). Such sensitiveness of the two methods, however, has not been fully assessed in my simulation due to the simple setting for the evolutionary process of molecular sequences. Therefore, for more practical evaluation for RY-coding and GG98 model, it is indispensable to re-assess their performance by analyzing real-world sequence dataset.

Here, I focused on a dataset used in Lau et al. (2009) [66], which comprises of protein-coding sequences 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 called as apicoplasts [67–69]. This dataset, henceforth designated as ‘Lau09 dataset,’ was used to infer the phylogenetic tree for investigating the origin of apicoplasts. The resultant tree topology strongly supported close relationship of apicoplasts with green algal or green alga-derived plastids [66], suggesting that apicoplasts were established through secondary endosymbiosis of a green alga; that is, the ‘green origin’ of

apicoplasts [70]. However, the green origin of apicoplasts supported by Lau et al 2009 [66] was contradictory to the widely accepted notion regarding apicoplasts as a residual endosymbiotic red alga; that is, the 'red origin' of apicoplasts [71–73].

I conjectured that above inconsistency in the origin of apicoplasts is attributed to the compositional bias in Lau09 dataset. Especially, the parallel shifts to extremely high AT content amongst sequences encoded in apicoplasts [74, 75] and particular green alga-derived plastids such as that of *Euglena longa* [76], can be the cause to mislead the phylogenetic inference into erroneously grouping them together. Importantly, extremely high AT content in protein-coding sequences in the plastid genomes of apicomplexa and *E. longa* can affect their codon usages, resulting in biased compositions of particular amino-acids coded by AT-rich codons; e.g., Phe, Ile, Lys, and Asn [74, 77, 78]. Thus, although the phylogenetic inferences in Lau et al. (2009) were based on amino-acid (AA) sequences and AA substitution models, they were still affected by the compositional bias stemming from the heterogeneity of AT content. Furthermore, extremely rapid substitution rates in apicoplasts and non-photosynthetic green-alga derived plastids, which were caused by their overall genome degeneration [74, 79], would make it more difficult to model substitution processes in Lau09 dataset.

For the above reasons, I regarded Lau09 dataset as a good example for the re-assessment for the performance of RY-coding and GG98 model under the presence of extraordinary AT content bias and complicated substitution processes in real-world sequences. In this study, I applied the above two approaches, as well as the homogeneous substitution model (HKY), to the ML analysis of the nucleotide format of the Lau09 dataset. Results obtained in this study clearly showed that RY-coding and GG98 model could recover the tree supporting the red origin of apicoplasts.

Nevertheless, I also observed that the statistical support for the close relationship of apicoplasts with red algal or red alga-derived plastids was clearly higher in the GG98 model-based analysis, compared with that in the RY-coding-based analysis.

III-2 Materials and Methods

III-2-1 Datasets

I 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 apicoplasts from GenBank database. The gene- and taxon-sampling in this study was a subset of Lau09 dataset [66]. For each gene, I firstly made a multiple alignment based on AA sequences by using MAFFT v.7 [80]. Resultant AA alignments were inspected by eye and manually edited. Then, the corresponding nt sequences were carefully aligned by referring their putative AA alignment using PAL2NAL [81]. After the exclusion of unambiguously aligned positions, the five single-gene nt alignments were concatenated into a '5-gene' alignment containing 31 taxa with 2,226 nt positions. Of note, the AT contents of sequences encoded in apicoplasts and green alga-derived plastid of *E. longa* are higher than other plastid sequences and produce significant compositional bias in the 5-gene alignment (Table 3 and Fig. 9A).

III-2-2 Tree comparison analysis

I firstly conducted the ML tree inference from the 5-gene alignment and the bootstrap analysis based on 100 replicates with the HKY + Γ model using PhyML v.3.0 [82]. The ML tree was selected from heuristic tree search based on the subtree pruning

and regrafting (SPR) method initiated from a parsimony tree. In the bootstrap analysis, a single tree search with SPR was performed per replicate. All parameters were estimated from the entire data. Consequently, the ML analysis placed five apicoplast sequences as the sister group to euglenids *(E. longa and E. gracilis*) within green algal/green alga-derived plastids (Fig. 9A), supporting the green origin of apicoplasts (left in Fig. 9B). Nevertheless, as mentioned above, this result is contradictory to the widely accepted hypothesis of the 'red origin' of apicoplasts (right in Fig. 9B). I anticipated that the tree topology representing the green origin of apicoplasts was attributed to the homogeneous (HKY) model ignoring the heterogeneity of AT content across taxa in the 5-gene alignment (Table. 3 and Fig. 9A). If the above conjecture was true, the ML analyses based on RY-coding and GG98 model would suppress the phylogenetic artifact and preferably select a tree topology representing the red origin of apicoplasts.

In order to examine the above conjecture, I prepared test trees representing the two competing hypotheses for the origin of apicoplasts by modifying the ML tree shown in Fig. 9A. The apicoplast clade was regrafted to i) seven terminal branches assuming the close relationship of apicoplasts to the single red/green algal species or red/green alga-derived plastid (highlighted by circles in Fig. 9A), and ii) seven internal blanches leading to a well-supported clade, assuming the affinity of apicoplasts to a certain group of alga or plastids (highlighted by diamonds in Fig. 9A). Subsequently, the ln*L*s for the 14 alternative trees (Fig. 10) were compared with that of the ML tree as described below.

The ln*L*s of the ML and 14 alternative trees were firstly calculated with the HKY + Γ (homogeneous) model using PhyML. The same ln*L* calculation was repeated with the GG98 + Γ model by using *eval_nh* program implemented in NHML v.3.0 [49,

53]. The root position was fixed in the second comparison with the GG98 + Γ model (highlighted by a star in Fig. 9A). Then, I recoded the original 5-gene-alignment by RY-coding and subjected recoded data to the ln*L* calculation of the ML and alternative trees based on the CF + Γ model using PhyML. In the above three analyses, branch lengths of all tree topologies were optimized and model parameters were estimated from the entire alignment.

III-2-3 Approximately unbiased test

Alternative positions of apicoplasts in the trees in Fig. 10 were examined by the approximately unbiased (AU) test [83] based on RY-coding and GG98 model. For Tree 0 through Tree 14, site-wise log-likelihoods (site-ln*L*s) were calculated based on the CF + Γ model with RY-recoded data using PhyML. The site-lnL data were then subjected to CONSEL v.0.2 with default parameter settings [84] in order to calculate the *p* value under the null hypothesis that the difference of the ln*L*s between the best tree and an alternative tree equals to 0. The same procedure was repeated based on the GG98 + Γ model.

III-2-4 ML tree search and bootstrap analysis based on RY-coding and GG98 model

In the ML analysis of RY-recoded data with the CF + Γ model, the ML tree was selected from heuristic tree search using SPR method. The tree search was initiated from Tree 9 in Fig. 10, which showed the highest lnL score in RY-coding-based tree comparison analysis. Whole model parameters were estimated from the data. After I obtained the ML tree, a bootstrap analysis was performed based on 100 bootstrap

replicates which were generated from original RY-recoded data. A single tree search with SPR was performed per replicate, starting from Tree 9 as mentioned above. I used PhyML for the ML tree search and the bootstrap analysis (MLBP analysis) based on RY-coding. The same process was repeated in the analysis with the GG98 + Γ model of the nt sequence data, except I added a new taxon, *Thermosynechococcus elongates*, which was retrieved from GenBank database as mentioned in III-2-1. The ML tree search from original data and each bootstrap replicate was initiated from Tree 9 in Fig. 10, where *T. elongates* was added as an out-group (this is necessary for the tree search with SPR method based on rooted tree). I used *shake_nh* program in NHML for the MLBP analysis based on the GG98 + Γ model.

III-3 Results

III-3-1 Results from tree comparison analysis

I subjected a real-world sequence dataset composed of five plastid-encoded genes, of which AT% varied from 56.2% to 84.59% amongst the taxa considered (Table 3 and Fig. 9A), to the ML analysis based on the homogeneous (HKY + Γ) model, as well as the GG98 + Γ model and the CF + Γ model (with RY-recoded data). The ML analysis of the 5-gene alignment with the HKY + Γ model, which cannot take into account the compositional heterogeneity (ΔAT%) across a tree, placed the apicoplast clade within green algal/green alga-derived plastids representing the green origin of apicoplasts (Fig. 9A). Then, I investigated whether the analyses with RY-coding and GG98 model suppresses the artifact from AT content heterogeneity in the 5-gene alignment (Table 3 and Fig. 9A), by assessing the position of apicoplasts. If both two approaches appropriately tolerate the compositional heterogeneity in the data, a tree

representing the red origin of apicoplasts should be preferred over those representing the alternative hypotheses including the green origin of apicoplasts.

I examined the origin of apicoplasts by comparing the ML tree inferred from the HKY model-based analysis (Fig. 9A) and 14 alternative trees, which are identical to the ML tree except for the position of apicoplasts (Fig. 10). In the tree comparison analysis based on the HKY+ Γ model, the ML tree (Tree 0; Fig. 10) received the highest ln*L* score among trees subjected to this comparison, preferring the artifactual green origin of apicoplasts. In contrast, both RY-coding-based analysis (RY analysis) and GG98 model-based analysis (GG98 analysis) supported the red origin of the apicoplast—Tree 9 in Fig. 10, in which the apicoplast clade grouped with red alga-derived plastids of diatoms (*Thalassiosira pseudonana* and *Odontella sinensis*), received higher ln*L* score than any other trees representing the green origin of apicoplasts (Fig. 10). These results indicate that ML phylogenetic analyses based on RY-coding and GG98 model successfully avoided a phylogenetic artifact stemming from AT content heterogeneity in the data. Nevertheless, I could observe no significant difference on ln*L* scores between Tree 9 and Tree 0 in both RY and GG98 analyses. The AU test failed to reject the null hypothesis of the ln*L*s being same between Tree 9 and Tree 0—the *p* value was 0.372 in RY analysis and 0.309 in GG98 analysis.

III-3-2 Results from MLBP analyses

In order to fully investigate the performance of RY-coding and GG98 model for reconstructing the accurate tree from the 5-gene-alignment, the ML tree search and bootstrap analysis with 100 replicates were performed for each method. Both RY-coding and GG98 analyses successfully placed the apicoplast clade within red algal/red

alga-derived plastids (Figs. 11A & 11B). However, I observed the difference on the bootstrap proportion (BP) value for the node uniting apicoplasts with red algal/red alga-derived plastids (highlighted by stars in Figs. 11A & 11B). The GG98 analysis supported the 'apicoplasts + red algal/red alga-derived plastids' clade with higher BP value (BP = 78; Fig. 11B) than the corresponding value from RY analysis (BP = 60; Fig. 11A). The support value discussed here directly reflects the phylogenetic signal uniting apicoplasts and red algal/red alga-derive plastids rather than green algal/green alga-derive plastids. Therefore, it can be proposed that the GG98 analysis exhibited better resolution for the red origin of apicoplasts than RY-coding analysis. Intriguingly, I also found the difference on the phylogenetic position of apicoplasts between RY-coding and GG98 analyses. In the ML tree inferred from RY-coding analysis, the apicoplast clade was placed as the sister to the clade of diatoms (Fig. 11A), while the clade was placed within red algae in the ML tree inferred from GG98 analysis (Fig. 11B). However, the grouping of apicoplasts neither with diatoms nor with red algae was supported by sufficiently high BP values (Figs 11A & 11B).

III-4 Discussion

Prior to this study, only a single study has applied both RY-coding and GG98 model to the ML analysis [85]. Husník et al. (2011) [86] showed that the two methods successfully suppressed the artifact that was strongly attracted by the AT content bias in a real-world sequence data. However, it was still ambiguous how efficiently RY-coding and GG98 model-based analyses reconstruct the accurate phylogenetic relationships from the data under study, because the 'true' phylogenetic tree is unknown in real-world data analyses. Contrary to Husník et al. 2011 [86], I here directly investigated the

performance of the analyses with RY-coding and GG98 model based on the position of apicoplasts in the tree for plastid-encoded sequences, assuming the 'true' and the 'false' hypotheses for the origin of apicoplasts.

In the 5-gene-alignment examined here, I observed remarkably high AT contents in sequences derived from five apicoplasts and the green alga-derived plastid in *Euglena longa* (AT = 73.9 ~ 84.6%; Table 3). The Student's *t*-test supported the average AT content of 78.1% among these 6 sequences was significantly higher than that calculated from other 25 sequences in the 5-gene alignment (average AT = 63.04%; $p = 5.38 \times 10^{-5}$). From the point of view of the red origin of apicoplasts [73], apicoplasts and the plastid in *E. longa* are distantly related to each other. However, as shown in my simulation, parallel shifts to extremely high AT% between distantly related lineages interrupt the accurate phylogenetic inference based on the homogeneous models. Indeed, the ML analysis using HKY model erroneously grouped these AT-rich sequences together and misled the artifactual green origin of apicoplasts (Fig. 9A). Of note, the topology of the ML tree shown in Fig 9A was consistent with the tree presented in Lau et al. (2009) [66], which was reconstructed from the amino-acid sequence data with a homogeneous amino-acid model. Thus, the compositional bias observed here cannot be mitigated by the translation from nucleotides (codons) to amino-acids, despite it has been considered to be an efficient approach to overcome the AT content bias in protein-coding sequence [86].

In contrast, both analyses with RY-coding and GG98 model properly selected the trees representing the red origin of apicoplasts (Figs. 10 and 11), demonstrating that these methods are robust enough against the compositional bias in the 5-gene-alignment. The maximum ΔAT% across taxa in the data reached to 28% (Table 3 and Fig. 9A).

Thus, a higher magnitude of the heterogeneity of AT content than that assumed in my simulation (see above) was tested in this analysis. Moreover, neither frequencies of A and T nor those of C and G were equal among all sequences in the 5-gene alignment (Table 3), representing the 'complex base composition' situation as assumed in my simulation (Fig. 7). In addition to such severe compositional bias, sequences derived from apicoplasts and plastid of parasitic green alga, *Helicosporidium* sp., exhibited extremely rapid substitution rates compared with any other sequences (Fig. 9A). This is due to the overall genome degeneration in these non-photosynthetic plastids [74, 79]. The rapid substitution rates would cause significant changes of substitution processes in these sequences, e.g., Ts/Tv ratio, and potentially affect the performance of the analyses of RY-coding (Fig. 4) and plausibly of GG98 model. Nonetheless, the results shown here revealed that both methods could retain their performance under the presence of the complicated evolutionary process of real-world sequences.

On the other hand, I also found that there was the difference on the resolution for the red origin of apicoplasts between the analyses with RY-coding and GG98 model. From the results of the MLBP analyses, GG98 model-based analysis showed superior performance compared to RY-coding-based analysis for detecting phylogenetic signal for the close relationship between apicoplasts and red alga/red algal-derived plastids (Figs. 11A and 11B). This might be attributed, at least to some extent, to erasing the true phylogenetic signal in the RY-coding analysis by recoding original sequence data. Therefore, I can conclude again that GG98 model is supposed to be more efficient than RY-coding for analyzing real-world sequence datasets.

IV. COMPUTATIONAL STUDY FOR ACCELERATING PHYLOGENETC INFERENCES BASED ON GG98 MODEL

IV-1 Introduction

The results obtained from the analyses of simulated and real-world sequence datasets support that GG98 model may be one of the most efficient approaches to ameliorate the ML phylogenetic inferences under the presence of strong AT content heterogeneity. On the other hand, the on-going accumulation of molecular sequence data driven by novel wet-lab techniques enables us to phylogenetically analyze large matrices composed of hundreds of genes derived from diverse organisms. Importantly, such 'large-scale phylogenetic analyses' can be significantly influenced by the heterogeneity of AT content across lineages [89–91], as large data size can enhance the artifactual impact of compositional heterogeneity in the homogeneous model-based analysis (Fig. 5). Therefore, it is strongly suggested that large-scale sequence datasets need to be analyzed by NH models. However, the phylogenetic inference based on GG98 model (and any other NH models) can be computationally much more intensive than those with homogeneous models. This is because the NH models require an enormous amount of model parameters to be optimized in a branch-by-branch fashion. In addition to this, the parameter optimization in the ML method involves the calculation of site-ln$L$ for each position, implying that the computational time for the analyses with NH models increases as more taxa and positions are included in our sequence data. Moreover, a comprehensive phylogenetic analysis (i.e., the ML tree search and bootstrap analysis) requires computing a lot of alternative trees. Consequently, the analysis based on GG98 or other NH models with large-scale

sequence data is beyond the capacity of a single CPU core on the personal computer systems.

On the other hand, recent advance in computational sciences has enabled us to run phylogenetic analyses using many CPU cores in parallel. To date, several pioneering works implemented efficient parallel computing methods in phylogenetic codes: OpenMP [92], MPI [93, 94], PTHREADS [95] and the combination of them [96]. These techniques were applied to parallelize various stages of the phylogenetic analysis, from the ln*L* calculation within a given tree to the computation of multiple trees during the ML tree search, and to the bootstrap analysis with multiple replicates. Nevertheless, all of currently available phylogenetic codes, which are applied to novel parallel computing techniques, only implement homogeneous models. Hence, it is urgent to develop a new program incorporating efficient parallel computing methods with NH models.

Here, I applied two parallel computing methods, OpenMP and MPI, to efficiently accelerate the calculation of ln*L*s across alternative trees based on GG98 model. The performance of the 'HYBRID' OpenMP/MPI code of NHML v3.0 [49] was benchmarked by analyzing simulated sequence datasets including ~130-taxon and ~10,000 nt positions. Consequently, I archived suitable speeding-up of the phylogenetic inference with the parallel version of NHML up to 64 computational nodes and 1,024 CPU cores on a supercomputer system, 'T2K-Tsukuba' (http://www.open-supercomputer.org/). This is de facto first computational effort to accelerate large-scale phylogenetic analyses with NH models.

IV-2 Materials and Methods

IV-2-1 Newton-Raphson (NR) algorithm in NHML

In the phylogenetic inference with NHML, the ln*L* score for a given tree is computed by optimizing branch lengths and model parameters of GG98 model [49]. Based on the ML method, the optimization of these parameters is operated by calculating their maximum-likelihood estimate (MLE) values. For this procedure, the Newton-Raphson (NR) algorithm [97] is implemented in NHML. The outline of NR algorithm is shown in Fig. 12.

In NR algorithm, the initial ln*L* score for a given tree is calculated according to Felsenstein (1981) [5]. Randomly determined values for branch lengths and model parameters of GG98 model are used in this initial ln*L* calculation (Fig. 12–(i)). Then, MLEs for parameters to be optimized (Θ) are computed by analytical method, in which the update of the ln*L* score and the values of Θ are iterated as defined in the *WHILE* loop in Fig. 12–(ii). In this iteration, first, the 1st and 2nd derivatives of the ln*L* with respect to each single parameter (θ) are respectively computed by fixing the values for any other parameters, as described in the first *FOR* loop in Fig. 12–(iii). Derivatives for θ are calculated from each site-ln*L* so that the calculation is repeated for the number of nt positions, as described in the second *FOR* loop (Fig. 12–(iv)). Second, each parameter θ gets updated based on the 2nd order Taylor approximation for the likelihood function (Fig. 12–(v)). Third, the ln*L* score for a given tree is re-calculated by reference to updated values for Θ (Fig. 12–(vi)), and the difference between current and previous ln*L*s (henceforth designated as Δln*L*) is calculated (Fig. 12–(vii)). As shown in the test condition for the *WHILE* loop (Fig. 12–(ii)), the procedure mentioned above will be iterated unless ΔlnL is less than sufficiently small value ε (in this study I used $\varepsilon = 0.1$).

When the iteration is finished, the ln*L* obtained from the final iteration step will be returned as the maximum ln*L* score for a given tree.

IV-2-2 Parallelization for NR algorithm

Upon the parallelization of NR algorithm, I focused on the two *FOR* loops for the calculation of derivatives (Fig. 12–(ii) and Fig. 12–(iii)). They occupy more than 90% of the total computational cost for NR algorithm due to the large number of iteration—the number of iteration increases in proportion to the number of taxa, which determines the number of parameters for branch lengths and branch-specific AT contents, and the number of nt positions. Therefore, the parallelization of these loops is most effective way to accelerate the calculation of ln*L* with NHML.

Here I applied two parallel computing methods, MPI and OpenMP, to the above two *FOR* loops. In this 'HYBRID' parallelization, the process-based parallelization by MPI was applied to the first *FOR* loop, while thread-based parallelization by OpenMP was applied to the second *FOR* loop. Thus, MPI processes are respectively assigned to optimize particular number of parameters in parallel, and each process controls multiple OpenMP threads to calculate site-wise derivatives for given parameters in parallel. Since each MPI process storages only the values of derivatives computed by itself, I applied the *MPI_Allgatherv* function to gather the data from each process and combine them into the complete data of derivatives for all parameters, which is then used for the update of parameters and ln*L* score (Fig. 12–(v) and Fig. 12–(iv)). In summary, the calculation of the ln*L* for a given tree based on *N* taxa and *M* positions is performed as described below.

1). The initial ln*L* is computed from randomly determined values for branch lengths and model parameters.

2). *P* MPI processes are respectively assigned to the parallel optimization of below parameters; 2*N*-3 parameters for branch lengths, 2*N*-2 for equilibrium AT contents on branches, one for ancestral AT content at the root, one for the root location, and one for the Ts/Tv ratio. Thus, the single MPI process optimizes (4*N*-2)/*P* parameters.

3). In each MPI process, *Q* OpenMP threads are respectively assigned to the parallel calculation of derivatives from site-ln*L*s of *M* positions. Thus, the single thread computes derivatives on *M*/*Q* positions.

4). Before updating parameters (Fig. 12–(v)), all MPI processes call *MPI_Allgatherv* function to gather the values of derivatives calculated in other processes, combine them, and broadcast the complete data of 1st and 2nd derivatives for all parameters to each other. Then, all parameters and ln*L* score are synchronously updated in each process.

5). Procedures 2)~4) will be iterated until the ln*L* score for a given tree would converge to the maximum value (Fig. 12–(ii)).

IV-2-3 Parallelization for the computation of multiple trees

The HYBRID parallelization for NR algorithm mentioned above is purposed to accelerate the calculation of ln*L* for a single tree. On the other hand, it is also considerable computational problem that we have to calculate ln*L*s for multiple trees during the ML tree search. Therefore, I here applied the method to efficiently distribute computational resources into the computation of multiple trees. As shown in Fig. 13, all

MPI processes are primarily controlled by the single MPI communicator called as '*MPI_COMM_WORLD*'. By dividing *MPI_COMM_WORLD* into several sub-communicators, which respectively control a partial group of MPI processes and OpenMP threads, I can assign them to the calculation of ln*L*s for different trees in parallel (Fig. 13). Of note, the ln*L* calculation for the single tree is also parallelized in each sub-communicator by the HYBRID code of NR algorithm (see IV-2-2).

IV-2-4 Benchmark datasets and experimental design

For the performance evaluation of parallelized NHML, I simulated nt sequence datasets based on 66-taxon- and 130-taxon-model trees. First, I prepare the 66-taxon model tree as shown in Fig. 14A. The lengths of branches leading to taxa 1 and 2, and taxa 63 and 64 (highlighted by red), were set to 1.0, while those of any other branches were set to 0.05. The ancestral sequence was randomly generated at the root (R in Fig. 14A), and each tip sequence was then simulated according to the given branch lengths. The substitution process was modeled by TN92 model [64], incorporating rate heterogeneity across sites approximated by a discrete gamma (Γ) distribution [58] with four categories. The κ parameter for Ts/Tv ratio [59] and the shape parameter α for a Γ distribution were set to 2.0 and 0.8. For the sequence simulation from the root to taxa 3–62, the AT content was set to 50%. On the other hand, sequences for taxa 1 and 2 and taxa 63 and 64 were designed to be AT-rich (AT = 90%), by changing the parameter for AT content in TN92 model at the node uniting these taxa (highlighted by red arrowheads in Fig. 14A). Under the above setting, I generated two datasets of different size, 2,500 nt positions and 10,000 nt positions (henceforth designated as 'small 66-taxon dataset' and 'large 66-taxon dataset').

Second, I prepared the 130-taxon model tree by bisecting the terminal branches on the 66-taxon model tree (diamonds in Fig. 14A). In this simulation, the lengths of branches leading to taxa 1 and 2 (generated by bisecting the branch leading to taxa 1 in Fig. 14A), and those leading to taxa 127 and 128 (generated by bisecting the branch leading to taxa 64 in Fig. 14A) were set to 1.0. On the other hand, any other branches were set to 0.05. All sequences were simulated following same model parameters as described above. Sequences of taxa 1 and 2 and taxa 127 and 128 evolved to be extremely AT-rich (AT = 90%) whereas those of any other taxa retained moderate level of AT content (AT = 50%). Consequently, I generated a sequence dataset of 2,500 nucleotide positions (henceforth designated as the '130-taxon dataset'). I used INDELible v.1.03 [98] for the sequence simulation.

Three simulation datasets were firstly subjected to the ML analyses based on the homogeneous GTR + Γ model [99] using RAxML v.8.0.0 [100]. Since the model cannot account the heterogeneity of AT content in the datasets into account, the artifactual trees which represent almost same topology as the model tree except erroneous grouping of AT-rich taxa were inferred from all datasets. For instance, taxa 63 and 64 were inferred to be wrongly united to the clade of taxa 1 and 2 in the analyses of both small and large 66-taxon datasets (Fig. 14B). Hence, I prepared alternative trees by changing the positions of taxa 63 and 64 in the ML tree. The clade of taxa 63 and 64, surrounded by red-broken line in Fig. 14B, was re-grafted to 16 terminal branches and 8 internal branches (stars in Fig. 14B) to generate 24 alternative trees. Note that these alternative trees include the 'true' tree, in which taxa 63 and 64 were placed as the sister group to taxa 61 and 62 (highlighted by a red star in Fig. 14B). Then, ln*L*s for 24 alternative trees were computed by parallelized NHML based on small and large

66-taxon datasets. Finally, I compared 24 alternative trees with the ML tree in Fig. 14B based on their ln*L*s re-calculated by GG98 + Γ model.

I also got the artifactual ML tree from the analyses of 130-taxon dataset with homogeneous GTR + Γ model, which exhibited the same topology as the 130-taxon model tree except that taxa 127 and 128 were erroneously grouped with taxa 1 and 2. Similar to the analyses of 66-taxon datasets, I prepared alternative trees by modifying the ML tree—the clade of taxa 127 and 128 were re-grafted to 32 terminal branches and 16 internal branches to generate 48 alternative trees, including the true tree in which these taxa were placed as the sister group to taxa 125 and 126. The ln*L* scores for the 48 alternative trees were also computed by parallelized NHML.

IV-2-5 Measurement environment

Computation of alternative trees based on 66-taxon and 130-taxon datasets was performed on T2K-Tsukuba supercomputer system (http://www.open-supercomputer.org/). The key characteristics of T2K-Tsukuba are listed in Table. 4. The single computational node on T2K-Tsukuba is composed of 4 sockets which respectively contain the quad-core CPU (AMD Opteron 8356, 2.30 GHz). In each node, one MPI process was assigned to one socket and 4 OpenMP threads, operated by the MPI process, were respectively allocated to 4 CPU cores. I used '*numactl –cpunodebind -localalloc*' options to conduct above HYBRID computing on T2K-Tsukuba. The total execution time for each benchmark run was measured by using ~64 computational nodes (i.e., ~1,024 CPU cores).

IV-3 Results

IV-3-1 Speeding-up by the HYBRID parallelization for NR algorithm

Benchmark runs for investigating the performance of the HYBRID code of NR algorithm were made for small and large 66-taxon datasets, and 130-taxon dataset. I used ~16 computational nodes (i.e., ~256 cores) of T2K-Tsukuba. Twenty-four and 48 alternative trees were computed respectively for 66-taxon- and 130-taxon-datasets. Of note, in all data analyses, I confirmed that the 'true' tree that is the same topology as the model tree was inferred to have the highest ln*L* score among all alternative trees. Based on the GG98 + Γ model, the ln*L* score for the true tree was higher than that for the artifactual tree, which was inferred from the homogeneous model-based analysis.

Changes of the total execution time for computing all alternative trees are shown in Fig. 15. The total execution time for the small 66-taxon dataset decreased approximately in reverse proportion to the number of CPU cores, and the benchmark run finally finished in 290 seconds on the use of 256 CPU cores (Fig. 15A). The same tendency was also observed in the analyses of the large 66-taxon dataset (Fig. 15B) and the 130-taxon dataset (Fig. 15C), where benchmark runs finally finished in 1,115 seconds and 1,150 seconds respectively. From the comparison of Figs. 15A and 15B, the scaling of the execution time was not significantly changed according to the number of nt positions, implying that OpenMP parallelization for the site-wise calculation of derivatives worked well regardless of the number of nt positions. Likewise, comparing Figs. 15B and 15C, it is also suggested that MPI parallelization successfully accelerated the parameter optimization regardless of the number of taxa. The results shown here indicate that the HYBRID parallelization largely improved the performance of the ln*L* calculation with NHML against variable scales of sequence dataset.

IV-3-2 Parallel efficiency of the HYBRID code of NR algorithm

The performance of the HYBRID code of NR algorithm was further inspected measuring the speeding-up ratios versus the number of CPU cores. As shown in Fig. 16, the HYBRID code showed good speeding-up up to 256 CPU cores in the analyses of all three sequence datasets. Nonetheless, the parallel efficiency (speeding-up per core) was gradually decreased as more CPU cores were used, and it finally dropped to 0.48 for small 66-taxon dataset, 0.56 for large 66-taxon dataset, and 0.65 for 130-taxon dataset (Fig. 16).

The drop in parallel efficiency was attributed to the overhead associated with *MPI_Allgatherv* communication, where each MPI process needs to gather the data of derivatives from other processes and then needs to broadcast the combined data to each other (see IV-2-2). In all three data analyses, the absolute time for *MPI_Allgatherv* communication was not significantly changed against the number of CPU cores (‘Comm time’ in Fig. 17), while the substantial time for the ln*L* calculation efficiently decreased (‘CPU time’ in Fig. 17). However, the occupancy of the Comm time in total execution time largely increased as more CPU cores (i.e., MPI processes) were used—it finally reached to 48.8% for small 66-taxon dataset (Fig. 17A), 43.1% for large 66-taxon dataset (Fig. 17B), and 35.7% for 130-taxon dataset (Fig. 17C). This is because the more processes are involved in *MPI_Allgatherv* communication, the larger the overhead for sending and receiving data between processes are incurred. Thus, the performance of the HYBRID code of NR algorithm was primarily restricted by the number of MPI processes assigned to the computation of the single tree.

Intriguingly, it was also revealed that speeding-up and parallel efficiency in the analyses of the large 66-taxon dataset and the 130-taxon dataset were significantly larger than those in the analysis of the small 66-taxon dataset (Fig. 16). The better performance for larger-scale sequence datasets is likely resulted from efficient reduction of CPU time over the increase in the Comm time (Fig. 17).

IV-3-3 Further speeding-up by the parallel computation of multiple trees

The performance of the parallel computation for multiple trees (see IV-2-3) was evaluated based on the 130-taxon dataset using ~1,024 CPU cores. I prepared three different schemes for partitioning MPI_COMM_WORLD into sub-communicators, which comprised of i) 16 CPU cores with 4 MPI processes, ii) 32 CPU cores with eight MPI processes, and iii) 64 CPU cores with 16 MPI processes. The ln*L* calculations for 48 alternative trees were equally distributed to each sub-communicator. See Table 5 for detailed numbers of trees computed by individual sub-communicators according to three partition schemes. Note that each MPI process operates 4 OpenMP threads and the calculation of the ln*L* for a single tree was performed by the HYBRID code of NR algorithm.

On the use of 256 CPU cores, I observed clear speeding-up for all three partition schemes compared to the control run, where all MPI processes are simultaneously assigned to compute a same tree without partitioning MPI_COMM_WORLD (Fig. 18). As shown in Fig. 17, the cost for *MPI_Allgatherv* communication becomes larger as more processes are assigned to the single ln*L* calculation. The cost for the MPI communication in each sub-communicator, thus, could be efficiently reduced by assigning relatively small number of MPI processes. On the

other hand, the total computational cost could also be decreased by the parallel computation of multiple trees using multiple, independently-working sub-communicators, resulting in the decrease of the total execution time under all partition schemes (Fig. 18).

Moreover, the schemes ii) and iii) showed further speeding-up up to 512 and 1,024 CPU cores respectively. Speeding-up ratios normalized by the run time on 256 CPU cores kept significantly high value: 1.75 for scheme ii) on 512 cores and 3.27 for scheme iii) on 1,024 cores. Finally, the parallel computing methods for multiple trees proposed here enabled the analysis of the 130-taxon dataset to finish in just 293 seconds on 1,024 CPU cores (Fig. 18), which is 40.1 times faster than using 16 CPU cores based on only the HYBRID parallelization of NR algorithm (Fig. 15C).

IV-4 Discussion

In this study, the phylogenetic inference with NHML was parallelized at multiple algorithmic levels. Fine- or medium-grained parallelization by OpenMP and MPI were applied to the calculation of the maximum ln*L* score for a given tree (IV-2-2), while coarse-grained parallelization by partitioning MPI_COMM_WORLD was applied to the computation of multiple trees (IV-2-3).

To date, the first and the third parallelisms mentioned above have been generally applied to phylogenetic inferences with homogeneous models [92–96, 101]. However, the second parallelism has not been emphasized due to relatively small number of parameters to be optimized in homogeneous model-based analyses. In contrast, phylogenetic inferences with GG98 model (and other NH models) potentially need to optimize piles of model parameters. Especially, the computational cost for

parameter optimization enormously increases when the sequence data of interest includes large number of taxa. Therefore, I here added a new MPI code for the parallel parameter optimization with the NH model, and combined it with an OpenMP code for the parallel computation of site-wise derivatives. As I expected, this HYBRID code showed good speeding-up against variable numbers of taxa and positions (Fig. 15 and 16), suggesting that both MPI and OpenMP parallelization worked well. It is also important to note that speeding-up and parallel efficiency were increased as more taxa and/or positions are included in the data (Fig. 16). Therefore, I can conclude that the HYBRID code of NHML can be well suited for the analyses of larger-scale sequence datasets.

Nevertheless, I also found that there was a limit of the HYBRID parallelization on the computation of a single tree. The parallel efficiency gradually decreased as more CPU cores were used in all three data analyses, and dropped to less than 0.5 in the analysis of the smallest dataset (Fig. 16). As a rule of thumb, parallelized codes are not able to work effective when the parallel efficiency becomes less than 0.5. The drop in parallel efficiency observed here was attributed to the rise of the cost for the communication among MPI processes (Fig. 17), implying that it’s not efficient to concentrate too much MPI processes at the computation of a single tree.

To keep efficient speeding-up on larger number of CPU cores and MPI processes, I here applied an upper level of parallelism in which MPI processes were partitioned into several small groups and allocated respectively to the computation of different trees. This coarse-grained parallelization showed further improvement for the speeding-up with various partition schemes (Fig. 18). Although I computed just 48 trees here, the parallelization proposed in this study can be expanded to compute larger

number of trees, by adjusting the size and the number of sub-communicators generated from MPI_COMM_WORLD. Thus, the heuristic ML tree search with NHML, which is performed by SPR method [49], can be efficiently parallelized by using this method. In conclusion, parallel version of NHML developed in the present study clearly showed well suited performance to accelerate ML phylogenetic inferences with GG98 model using more than one thousand CPU cores on a current high-performance computer system.

The HYBRID parallel computing methods proposed here can be applied to more flexible NH models [48], as they use NR algorithm for the ln$L$ calculation. Moreover, the bootstrap analysis with NH models, which requires the highest computational cost, can also accelerated by expanding the partitioning schemes of MPI processes [93, 96], or by adding new parallel computing methods such as GPGPU computing [101–103] or many-core computing [104], to compute multiple bootstrap replicates in parallel. Finally, the computational effort demonstrated in this study can lay a base for future works to establish fast and accurate phylogenetic codes toward large-scale comprehensive phylogenetic analyses based on NH models.

V. GENERAL DISCUSSION

V-1 Pros and Cons of RY-coding and GG98 model

RY-coding and GG98 model are underpinned by two different concepts for overcoming phylogenetic artifacts stemming from AT content heterogeneity—the former method aims to homogenize the compositional heterogeneity in sequence data by character recoding, while the latter method focuses on theoretically describing the non-homogeneous sequence evolution across a tree [49].

Several experimental studies, including the present simulation and real-world analyses, have suggested that both two methods can efficiently ameliorate the ML analyses compared to the conventional method using homogeneous substitution models [44, 60, 51, 52, 89, 53, 54]. However, my comprehensive survey revealed that there are innegligible differences on the performance between RY-coding and GG98 model-based phylogenetic analyses. RY-coding can greatly improve the accuracy of tree inference in spite of its simplicity, i.e., the artifactual impact of the heterogeneity of AT content can be greatly mitigated by just recoding original sequence data. Nevertheless, we have to pay attention to the potential pitfalls of this method. First, the robustness of RY-coding-based analysis, at least to some extent, depends on the substitution process that generated the data of interest (e.g., Ts/Tv ratio). Furthermore, if compositional heterogeneity in the data cannot be completely homogenized by cahracter recoding, RY-coding-based analysis may mislead to the artifact (Chapter II). Second, the resolution for the true phylogenetic relationships in RY-coding-based analysis may be decreased compared to the analysis based on GG98 model because the true phylogenetic signal in the original sequence data can be erased by recoding (Chapter III).

On the other hand, GG98 model is supposed to be free from the above issues in RY-coding. GG98 model can tolerate various degree of Ts/Tv ratio (Chapter II) and detect much more phylogenetic signal by analyzing original nt sequences (Chapter III). Although the robustness of GG98 model may be significantly depressed in case that the model assumption would be violated by complex compositional heterogeneity, we can resolve this problem by adding appropriate model parameters (Chapter II). Such flexibility of GG98 model (and other NH models) can be a strong evidence to suggest them as the most robust, and the most powerful approaches to reconstruct accurate phylogenetic trees from real-world sequence datasets bearing various degrees of compositional heterogeneity. However, the ML analyses based on GG98 model, as well as other NH models, can be computationally intense due to an enormous number of model parameters to be optimized, whereas we just need to optimize much less parameters by analyzing the simple binary data in RY-coding-based analyses. The computational time for the analyses with NH models can be reduced by implementing parallel computing methods as demonstrated in this study (Chapter IV), albeit we must require many computational resources (CPU cores) for the fast phylogenetic inference from large-scale sequence dataset.

Hence, we should be aware that there is a trade-off of the computational cost and the performance of phylogenetic inference between the data-recoding method and NH models. Present study strongly emphasizes the importance of using either or both of these methodologies properly according to the sequence data of interest.

V-2 How to infer accurate phylogenetic trees from sequence data with extraordinary compositional heterogeneity

In conclusion, I here propose a guideline for inferring the accurate phylogenetic tree under the presence of extraordinary base compositional bias, considering the merits and demerits of the data-recoding method and NH models discussed above.

If the heterogeneity of base composition exists, the nt sequence data should principally be subjected to RY-coding in order to check whether the compositional heterogeneity, especially the heterogeneity of AT content, can be successfully homogenized by RY-coding. If so, then we can subject the recorded data to the ML tree search and the bootstrap analysis (MLBP analysis) with the CF model, followed by the comparison of inferred tree topology and corresponding BP values with those obtained from the original nt data using homogeneous nt models. We might see the significant change of tree topology including i) collapse of the relationships between taxa which were erroneously grouped due to the compositional heterogeneity, and ii) reposition of those taxa into potentially accurate phylogenetic positions. However, I strongly recommend running the MLBP analysis based on GG98 model as well, in case that the performance of RY-coding would be influenced by the substitution process in the data and/or by the loss of important phylogenetic information by recoding. The HYBRID version of NHML developed here can be utilized for this procedure. From the results, we may see the consistence or inconsistence of the tree topology and its BP support values between RY-coding and GG98 model-based analyses. Unfortunately, currently no method is available for comparing the appropriateness between the data-recoding method and NH models based on statistical criteria like AIC [105] or BIC [106], because the data analyzed in these methods are not identical (i.e., binary data and nt

sequence data). Therefore, it is necessary to develop a new procedure to statistically compare the two methods under different data types.

I have to anticipate the case that a quite complicated heterogeneity of base composition is exhibited by the sequence data of interest; thus, the compositions of four bases are not similar within sequence, and even among sequences. In such case, any data-recoding methods that convert four bases into two-state characters would not work well because they are not capable of homogenizing the compositional heterogeneity in the data. It is strongly recommended to apply an appropriate NH model for describing the substitution process that generated the observed compositional heterogeneity among taxa. At the same time, statistical selection of the most appropriate NH model for the data of interest is necessary to avoid over-fitting (over-parameterization) which causes increase of computational time and statistical error for parameter optimization. Hence, advanced programs for model selection, such as *testnh* program implemented in Bio++ package [107], are indispensable before applying NH models to phylogenetic analyses.

It is also supposed to be efficient to use amino-acid data for protein-coding sequences, as the translation from nucleotides to amino-acids can cancel the compositional bias at 3rd codon positions [86]. Importantly, the data-recoding method and NH models can be applied to the analyses of amino-acid sequences in a similar way with nt sequences [48, 108]. It enables us to expect that the ML phylogenetic analyses based on amino-acid sequences which are still suffered from compositional heterogeneity even after the translation, as seen in the data used in Lau et al. (2009) [66], would be improved by applying the data-recoding method and NH models. However, the basic properties of the above methods are still unknown since no simulation studies, as well as experimental real-world analyses, are currently available. Furthermore, NH

models for amino-acid sequences are significantly difficult to compute due to a pile of model parameters for describing amino-acid substitution process. Finally, future assessment and computational challenge would help us to advance our knowledge and techniques for inferring the most accurate phylogenetic trees from diverse empirical sequence datasets, which bear a variety of compositional heterogeneity.

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Dr. Tetsuo Hashimoto (University of Tsukuba) for insightful comments and warm encouragement during the course of my study. I am grateful to Dr. Yuji Inagaki (University of Tsukuba), Dr. Mitsuhisa Sato (University of Tsukuba), and Dr. Ken-ichiro Ishida (University of Tsukuba) for their critical reading of the manuscript. I thank to Dr. Masahiro Nakao (RIKEN Advanced Institute for Computational Science) for technical guidance on my computational study. I also thank all members of the laboratory of Molecular Evolution of Microbes, in University of Tsukuba.

I am supported by the JSPS Research Fellowship for Young Scientists DC1 (Nos. 24007).

References

1. Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology**. *Syst Biol* 1971, **20**:406–416.

2. Felsenstein J: **Statistical inference of phylogenies.** *J R Stat Soc Ser A* 1983, **146**:246–272.

3. Sokal R, Michener C: **A statistical method for evaluating systematic relationships**. *Univ Kans Sci Bull* 1958, **38**:1409 – 1438.

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

5. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17**:368–76.

6. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading**. *Syst Biol* 1978, **27**:401–410.

7. Anderson FE, Swofford DL: **Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA.** *Mol Phylogenet Evol* 2004, **33**:440–51.

8. Philippe H: **Opinion: long branch attraction and protist phylogeny.** *Protist* 2000, **151**:307–316.

9. Sanderson MJ, Wojciechowski MF, Hu J-M, Khan TS, Brady SG: **Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants**. *Mol Biol Evol* 2000, **17**:782–797.

10. 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**:843–849.

11. Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS: **Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods.** *Syst Biol* 2001, **50**:525–539.

12. Inagaki Y, Susko E, Fast NM, Roger AJ: **Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF-1 α phylogenies.** *Mol Biol Evol* 2004, **21**:1340–1349.

13. Philippe H, Germot A: **Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution.** *Mol Biol Evol* 2000, **17**:830–834.

14. Kelchner SA, Thomas MA: **Model use in phylogenetics: nine key questions.** *Trends Ecol Evol* 2007, **22**:87–94.

15. Foster PG, Hickey D a: **Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions.** *J Mol Evol* 1999, **48**:284–290.

16. Mooers A, Holmes E: **The evolution of base composition and phylogenetic inference.** *Trends Ecol Evol* 2000, **15**:365–369.

17. Felsenstein J: **Phylogenies from molecular sequences: inference and reliability.** *Annu Rev Genet* 1988, **22**:521–565.

18. Rosenberg MS, Kumar S: **Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference.** *Mol Biol Evol* 2003, **20**:610–621.

19. Galtier N, Lobry JR: **Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes.** *J Mol Evol* 1997, **44**:632–636.

20. Haywood-Farmer E, Otto SP: **The evolution of genomic base composition in bacteria.** *Evolution* 2003, **57**:1783–1792.

21. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M: **The 160-kilobase genome of the bacterial endosymbiont *Carsonella*.** *Science* 2006, **314**:267.

22. Karlin S, Mrázek J: **Compositional differences within and between eukaryotic genomes.** *Proc Natl Acad Sci U S A* 1997, **94**:10227–10232.

23. Clark CG, Alsmark UCM, Tazreiter M, Saito-Nakano Y, Ali V, Marion S, Weber C, Mukherjee C, Bruchhaus I, Tannich E, Leippe M, Sicheritz-Ponten T, Foster PG, Samuelson J, Noël CJ, Hirt RP, Embley TM, Gilchrist CA, Mann BJ, Singh U, Ackers JP, Bhattacharya S, Bhattacharya A, Lohia A, Guillén N, Duchêne M, Nozaki T, Hall N: **Structure and content of the *Entamoeba histolytica* genome.** *Adv Parasitol* 2007, **65**:51–190.

24. Dávalos LM, Perkins SL: **Saturation and base composition bias explain phylogenomic conflict in *Plasmodium*.** *Genomics* 2008, **91**:433–442.

25. Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG: **From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes.** *BMC Evol Biol* 2014, **14**:23.

26. Carulli JP, Krane DE, Hartl DL, Ochman H: **Compositional heterogeneity and patterns of molecular evolution in the *Drosophila* genome.** *Genetics* 1993, **134**:837–845.

27. Romiguier J, Ranwez V, Douzery EJP, Galtier N: **Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes.** *Genome Res* 2010, **20**:1001–1009.

28. Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A: **Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system**. *Gene* 1999, **238**:195–209.

29. Smith DR, Lee RW: **Mitochondrial genome of the colorless green alga *Polytomella* capuana: a linear molecule with an unprecedented GC content.** *Mol Biol Evol* 2008, **25**:487–496.

30. Smith DR: **Unparalleled GC content in the plastid DNA of *Selaginella*.** *Plant Mol Biol* 2009, **71**:627–639.

31. Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James TY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Standridge SE, Nerad TA, Shearer CA, Smirnov A V, Spiegel FW, Taylor MFJR: **The new higher level classification of eukaryotes with emphasis on the taxonomy of protists.** *J Eukaryot Microbiol* 2005, **52**:399–451.

32. Rodríguez-Trelles F, Tarrío R, Ayala FJ: **Fluctuating mutation bias and the evolution of base composition in *Drosophila*.** *J Mol Evol* 2000, **50**:1–10.

33. Hasegawa M, Hashimoto T: **Ribosomal RNA trees misleading?** *Nature* 1993, **361**:23.

34. Chang BS, Campbell DL: **Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences.** *Mol Biol Evol* 2000, **17**:1220–1231.

35. Tarrío 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**:1464–1473.

36. Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW: **The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated.** *Syst Biol* 2004, **53**:638–643.

37. Ho SY, Jermiin L: **Tracing the decay of the historical signal in biological sequence data.** *Syst Biol* 2004, **53**:623–637.

38. Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF: **When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics**. *Syst Entomol* 2010, **35**:429–448.

39. Nesnidal MP, Helmkampf M, Bruchhaus I, Hausdorf B: **Compositional heterogeneity and phylogenomic inference of metazoan relationships.** *Mol Biol Evol* 2010, **27**:2095–2104.

40. Groussin M, Gouy M: **Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea.** *Mol Biol Evol* 2011, **28**:2661–2674.

41. Sheffield NC, Song H, Cameron SL, Whiting MF: **Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics.** *Syst Biol* 2009, **58**:381–394.

42. Cavender JA, Felsenstein J: **Invariants of phylogenies in a simple case with discrete states**. *J Classif* 1987, **4**:57–71.

43. Phillips MJ, Penny D: **The root of the mammalian tree inferred from whole mitochondrial genomes**. *Mol Phylogenet Evol* 2003, **28**:171–185.

44. Phillips MJ, Delsuc F, Penny D: **Genome-scale phylogeny and the detection of systematic biases.** *Mol Biol Evol* 2004, **21**:1455–1458.

45. Yang Z, Roberts D: **On the use of nucleic acid sequences to infer early branchings in the tree of life.** *Mol Biol Evol* 1995, **12**:451–458.

46. Foster PG: **Modeling compositional heterogeneity.** *Syst Biol* 2004, **53**:485–495.

47. Jayaswal V, Jermiin LS, Poladian L, Robinson J: **Two stationary nonhomogeneous Markov models of nucleotide sequence evolution.** *Syst Biol* 2011, **60**:74–86.

48. Dutheil J, Boussau B: **Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs.** *BMC Evol Biol* 2008, **8**:255.

49. 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**:871–879.

50. Galtier N, Gouy M: **Inferring phylogenies from DNA sequences of unequal base compositions.** *Proc Natl Acad Sci U S A* 1995, **92**:11317–11321.

51. Phillips MJ, Lin YH, Harrison GL, Penny D: **Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials.** *Proc Biol Sci* 2001, **268**:1533–1538.

52. Nabholz B, Künstner A, Wang R, Jarvis ED, Ellegren H: **Dynamic evolution of base composition: causes and consequences in avian phylogenomics.** *Mol Biol Evol* 2011, **28**:2197–2210.

53. Galtier N, Tourasse N, Gouy M: **A nonhyperthermophilic common ancestor to extant life forms.** *Science* 1999, **283**:220–221.

54. Herbeck JT, Degnan PH, Wernegreen JJ: **Nonhomogeneous model of sequence evolution indicates independent origins of primary endosymbionts within the enterobacteriales (gamma-Proteobacteria).** *Mol Biol Evol* 2005, **22**:520–532.

55. Masta SE, Longhorn SJ, Boore JL: **Arachnid relationships based on mitochondrial genomes: asymmetric nucleotide and amino acid bias affects phylogenetic analyses.** *Mol Phylogenet Evol* 2009, **50**:117–128.

56. 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**:2581–2593.

57. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA**. *J Mol Evol* 1985, **22**:160–174.

58. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses**. *Trends Ecol Evol* 1996, **11**:367–372.

59. Yang Z, Yoder AD: **Estimation of the transition/transversion rate bias and species Sampling**. *J Mol Evol* 1999, **48**:274–283.

60. Phillips MJ, Penny D: **The root of the mammalian tree inferred from whole mitochondrial genomes**. *Mol Phylogenet Evol* 2003, **28**:171–185.

61. Swofford D: **PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.** 2003.

62. Harshman J, Braun EL, Braun MJ, Huddleston CJ, Bowie RCK, Chojnowski JL, Hackett SJ, Han K-L, Kimball RT, Marks BD, Miglia KJ, Moore WS, Reddy S, Sheldon FH, Steadman DW, Steppan SJ, Witt CC, Yuri T: **Phylogenomic evidence for multiple losses of flight in ratite birds.** *Proc Natl Acad Sci U S A* 2008, **105**:13462–13467.

63. Gruber KF, Voss RS, Jansa S a: **Base-compositional heterogeneity in the RAG1 locus among didelphid marsupials: implications for phylogenetic inference and the evolution of GC content.** *Syst Biol* 2007, **56**:83–96.

64. 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**:678–687.

65. Keller I, Bensasson D, Nichols RA: **Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes.** *PLoS Genet* 2007, **3**:e22.

66. Lau AOT, McElwain TF, Brayton K a, Knowles DP, Roalson EH: ***Babesia bovis*: a comprehensive phylogenetic analysis of plastid-encoded genes supports green algal origin of apicoplasts.** *Exp Parasitol* 2009, **123**:236–243.

67. Waller RF, McFadden GI: **The apicoplast: a review of the derived plastid of apicomplexan parasites.** *Curr Issues Mol Biol* 2005, **7**:57–79.

68. Maréchal E, Cesbron-Delauw M-F: **The apicoplast: a new member of the plastid family**. *Trends Plant Sci* 2001, **6**:200–205.

69. Foth BJ, McFadden GI: **The apicoplast: A plastid in *Plasmodium falciparum* and other apicomplexan parasites**. *Int Rev Cytol* 2003, **224**:57–110.

70. Funes S, Davidson E, Reyes-Prieto A, Magallón S, Herion P, King MP, González-Halphen D: **A green algal apicoplast ancestor.** *Science* 2002, **298**:2155.

71. Blanchard JL, Hicks JS: **The non-photosynthetic plastid in malarial parasites and other apicomplexans is derived from outside the green plastid lineage**. *J Eukaryot Microbiol* 1999, **46**:367–375.

72. Williamson DH, Gardner MJ, Preiser P, Moore DJ, Rangachari K, Wilson RJM: **The evolutionary origin of the 35 kb circular DNA of *Plasmodium falciparum*: new evidence supports a possible rhodophyte ancestry**. *Mol Gen Genet MGG* 1994, **243**:249–252.

73. Janouskovec J, Horák A, Oborník M, Lukes J, Keeling PJ: **A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids.** *Proc Natl Acad Sci U S A* 2010, **107**:10949–10954.

74. Wilson RJ, Denny PW, Preiser PR, Rangachari K, Roberts K, Roy a, Whyte a, Strath M, Moore DJ, Moore PW, Williamson DH: **Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*.** *J Mol Biol* 1996, **261**:155–172.

75. Cai X, Fuller AL, McDougald LR, Zhu G: **Apicoplast genome of the coccidian *Eimeria tenella*.** *Gene* 2003, **321**:39–46.

76. Gockel G, Hachtel W: **Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*.** *Protist* 2000, **151**:347–351.

77. Foster PG, Jermiin LS, Hickey DA: **Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria**. *J Mol Evol* 1997, **44**:282–288.

78. Singer G a, Hickey D a: **Nucleotide bias causes a genomewide bias in the amino acid composition of proteins.** *Mol Biol Evol* 2000, **17**:1581–1588.

79. De Koning AP, Keeling PJ: **The complete plastid genome sequence of the parasitic green alga *Helicosporidium* sp. is highly reduced and structured.** *BMC Biol* 2006, **4**:12.

80. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30**:772–780.

81. Suyama M, Torrents D, Bork P: **PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.** *Nucleic Acids Res* 2006, **34**(Web Server issue):W609–612.

82. Guindon S, Dufayard J-F, 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**:307–321.

83. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst Biol* 2002, **51**:492–508.

84. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection**. *Bioinformatics* 2001, **17**:1246–1247.

85. Husník F, Chrudimský T, Hypša V: **Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches.** *BMC Biol* 2011, **9**:87.

86. Hashimoto T, Nakamura Y, Nakamura F, Shirakura T, Adachi J, Goto N, Okamoto K, Hasegawa M: **Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*.** *Mol Biol Evol* 1994, **11**:65–71.

87. Fast NM, Kissinger JC, Roos DS, Keeling PJ: **Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids**. *Mol Biol Evol* 2001, **18**:418–426.

88. Oborník M, Janouskovec J, Chrudimský T, Lukes J: **Evolution of the apicoplast and its hosts: from heterotrophy to autotrophy and back again.** *Int J Parasitol* 2009, **39**:1–12.

89. Sheffield NC, Song H, Cameron SL, Whiting MF: **Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics.** *Syst Biol* 2009, **58**:381–394.

90. Jeffroy O, Brinkmann H, Delsuc F, Philippe H: **Phylogenomics: the beginning of incongruence?** *Trends Genet* 2006, **22**:225–231.

91. Li B, Lopes JS, Foster PG, Embley TM, Cox CJ: **Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins.** *Mol Biol Evol* 2014, **31**:1697–1709.

92. Stamatakis A, Ott M, Ludwig T: **RAxML-OMP : An efficient program for phylogenetic inference on SMPs**. *Parallel Comput Technol* 2005, **3606**:288–302.

93. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**:2688–2690.

94. Stamatakis A, Aberer AJ: **Novel parallelization schemes for large-scale likelihood-based phylogenetic inference**. In *2013 IEEE 27th International Symposium on Parallel and Distributed Processing*. IEEE; 2013:1195–1204.

95. Stamatakis A, Ott M: **Exploiting fine-grained parallelism in the phylogenetic likelihood function with MPI, Pthreads, and OpenMP: a performance study**. *Pattern Recognit Bioinforma* 2008:424–435.

96. Pfeiffer W, Stamatakis A: **Hybrid MPI/Pthreads parallelization of the RAxML phylogenetics code**. *2010 IEEE Int Symp Parallel Distrib Process Work Phd Forum* 2010:1–8.

97. Felsenstein J, Churchill GA: **A Hidden Markov Model approach to variation among sites in rate of evolution**. *Mol Biol Evol* 1996, **13**:93–104.

98. Fletcher W, Yang Z: **INDELible: a flexible simulator of biological sequence evolution.** *Mol Biol Evol* 2009, **26**:1879–1888.

99. Lanave C, Preparata G, Sacone C, Serio G: **A new method for calculating evolutionary substitution rates**. *J Mol Evol* 1984, **20**:86–93.

100. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30**:1312–1313.

101. Pratas F, Trancoso P, Stamatakis A, Sousa L: **Fine-grain parallelism using multi-core, Cell/BE, and GPU Systems: Accelerating the phylogenetic likelihood function**. *2009 Int Conf Parallel Process* 2009:9–17.

102. Izquierdo-Carrasco F, Alachiotis N, Berger S, Flouri T, Pissis SP, Stamatakis A: **A generic vectorization scheme and a GPU kernel for the phylogenetic likelihood library**. In *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*. IEEE; 2013:530–538.

103. Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, Rambaut A, Suchard M a: **BEAGLE: an**

application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol* 2012, **61**:170–173.

104. Kozlov AM, Goll C, Stamatakis A: **Efficient computation of the phylogenetic likelihood function on the intel MIC architecture**. In *Proceedings of the 2014 IEEE International Parallel & Distributed Processing Symposium Workshops (IPDPSW '14)*. IEEE Computer Society; 2014:518–527.

105. Akaike H: **A new look at the statistical model identification**. *IEEE Trans Automat Contr* 1974, **19**:716–723.

106. Schwarz G: **Estimating the dimension of a model**. *Ann Stat* 1978, **6**:461–464.

107. Dutheil JY, Galtier N, Romiguier J, Douzery EJP, Ranwez V, Boussau B: **Efficient selection of branch-specific models of sequence evolution.** *Mol Biol Evol* 2012, **29**:1861–1874.

108. Foster PG, Cox CJ, Embley TM: **The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods.** *Philos Trans R Soc Lond B Biol Sci* 2009, **364**:2197–2207.

TABLES

Table 1. Settings for the base frequencies applied to the terminal branches leading to Taxa 3 and 4 in the sequence generation (1st, 2nd, 4th, and 5th columns), and the average AT content (AT%) in the resultant Taxa 3 and 4 sequences (3rd and 6th columns).

Ts/Tv ratio (κ) = 2.0			Ts/Tv ratio (κ) = 0.2		
Settings of base frequencies in data simulation (%)		Average AT% achieved in 500 replicates (%) (mean ± 2*SD)	Settings of base frequencies in data simulation (%)		Average AT% achieved in 500 replicates (%) (mean ± 2*SD)
A & T	G & C		A & T	G & 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

Table 2. Settings for the base frequencies applied to the terminal branches leading to Taxa 3 and 4 in sequence generation (1st, 2nd, 4th, and 5th columns), and the average value of the difference of purine (R) between Taxa 1 and 2, and Taxa 3 and 4 in RY-recoded sequences (ΔR%; 3rd and 6th columns).

Ts/Tv ratio (κ) = 2.0			Ts/Tv ratio (κ) = 0.2		
Settings of base frequencies in data simulation (%)		Average ΔR in 500 replicates (%) (mean ± 2*SD)	Settings of base frequencies in data simulation (%)		Average ΔR in 500 replicates (%) (mean ± 2*SD)
A & T	G & C		A & T	G & C	
25.0	25.0	1.99 ± 0.6	25.0	25.0	1.99 ± 0.6
26.5	23.5	1.99 ± 0.6	27.0	23.0	2.01 ± 0.7
28.0	22.0	1.99 ± 0.6	29.0	21.0	1.99 ± 0.6
29.5	20.5	2.00 ± 0.6	31.0	19.0	1.99 ± 0.6
31.0	19.0	1.99 ± 0.6	33.0	17.0	1.99 ± 0.6
32.5	17.5	2.00 ± 0.6	35.0	15.0	1.99 ± 0.6
34.0	16.0	2.01 ± 0.6	37.0	13.0	1.98 ± 0.6
35.5	14.5	2.02 ± 0.6	39.0	11.0	1.97 ± 0.6
37.0	13.0	2.00 ± 0.6	41.0	9.0	1.98 ± 0.6
38.5	11.5	2.00 ± 0.6	43.0	7.0	1.97 ± 0.6
40.0	10.0	2.00 ± 0.7	45.0	5.0	1.98 ± 0.6

Table 3. The heterogeneity of base composition and AT content across taxa in the 5-gene alignment.

Taxon name	A (%)	T (%)	G (%)	C (%)	A+T (%)
Babesia bovis	39.71	34.14	13.97	12.17	73.85
Theileria parva	41.87	37.69	11.86	8.58	79.56
Plasmodium falciparum	45.24	39.35	9.88	5.53	84.59
Eimeria tenella	42.36	34.59	13.16	9.88	76.95
Toxoplasma gondii	40.70	38.81	12.40	8.09	79.52
Eunglena longa	42.86	31.04	15.14	10.96	73.90
Euglena gracilis	33.29	33.60	20.35	12.76	66.89
Oryza nivara	31.72	27.90	22.87	17.52	59.61
Arabidopsis thaliana	31.45	28.48	23.14	16.94	59.93
Anthoceros formosae	32.12	29.70	22.42	15.77	61.82
Chaetosphaeridium globosum	34.82	31.45	19.86	13.88	66.26
Mesostigma viride	33.29	31.13	20.85	14.74	64.42
Chlorella vulgaris	29.96	30.01	21.61	18.42	59.97
Helicosporidium sp.	35.27	34.41	16.85	13.48	69.68
Bigelowiella natans	34.41	33.92	18.87	12.80	68.33
Pseudenoclonium akinetum	30.86	31.85	21.29	15.99	62.71
Oltmannsiellopsis viridis	30.14	30.77	21.11	17.97	60.92
Scenedesmus obliqus	32.44	33.29	19.68	14.60	65.72
Chlamydomonas reinhardtii	31.00	33.06	20.89	15.05	64.06

Table 3. The heterogeneity of base composition and AT content across taxa in the 5-gene alignment.

Stigeoclonium helvetiucum	32.08	31.40	21.79	14.74	63.48
Leptosia terrestris	32.84	31.76	20.40	15.00	64.60
Nephroselmis olivacea	27.94	28.26	23.59	20.22	56.20
Thalassiosira pseudonana	33.74	31.00	19.90	15.36	64.74
Odontella sinensis	31.72	30.77	20.98	16.53	62.49
Rhodomonas salina	33.65	27.22	22.15	16.98	60.87
Guillardia theta	33.96	29.16	20.71	16.17	63.12
Cyanidium caldarium	33.51	29.34	21.56	15.59	62.85
Cyanidioschyzon merrolae	29.96	30.55	22.87	16.62	60.51
Porphyra purpurea	31.99	29.38	22.24	16.40	61.37
Gracilaria tennuistipitata	35.09	29.20	20.76	14.96	64.29
Emiliania huxleyi	31.81	29.52	21.79	16.89	61.32

Table 4. Specification of the performance measurement environment on T2K-Tsukuba supercomputer system.

CPU	Quad-core AMD Opteron 8356 (2.30GHz) per socket (4 sockets / node)
Memory	DDR2, 667MHz, 2GB x 16 = 32GB per node
Network	Infiniband 4xDDR, Mellanox ConectX x 4
Compiler	GCC 4.6.4
MPI Library	MVAPICH2 v.1.7

Table 5. Number of trees computed in each sub-communicator in each partition scheme. 48 alternative trees to be computed for 130-taxon dataset were equally distributed to sub-communicators, where 4, 8, and 16 MPI processes were respectively allocated according to three different partition schemes for MPI_COMM_WORLD.

	Number of CPU cores		
Partition schemes	256	512	1024
4 processes / sub-comm	3		
8 processes / sub-comm	6	3	
16 processes / sub-comm	12	6	3

FIGURES

Fig. 1. A Four-taxon tree for showing the calculation of likelihood. The tree is composed of four external branches, t_1–t_4, and one internal branches, t_5. Bases observed at extant taxa, taxon 1–4, are represented as *p*, *q*, *r*, *s*, while those at internal nodes are defined as *i* and *j*.

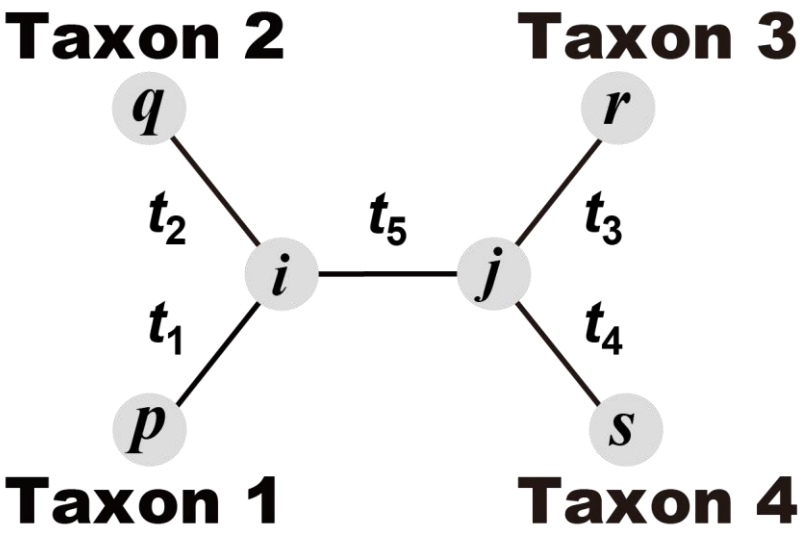

Taxon 2
Taxon 3
q
r
t2
t5
t3
i
j
t1
t4
p
s
Taxon 1
Taxon 4

Fig. 2. 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 (Γ) distribution and Ts/Tv ratio were fixed across a tree. The frequencies for A, C, G, and T were set to 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.

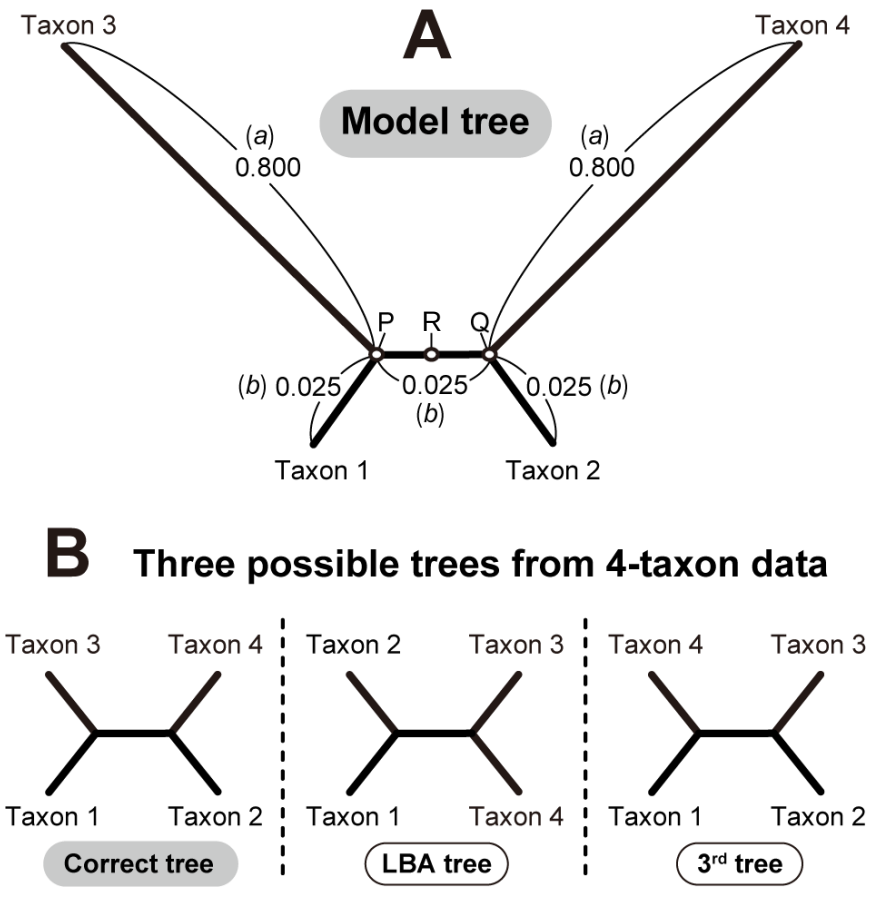

A
Model tree
Taxon 3
Taxon 4
(a)
0.800
(a)
0.800
P
R
Q
(b) 0.025
0.025
(b)
0.025 (b)
Taxon 1
Taxon 2
B
Three possible trees from 4-taxon data
Taxon 3
Taxon 4
Taxon 1
Taxon 2
Correct tree
Taxon 2
Taxon 3
Taxon 1
Taxon 4
LBA tree
Taxon 4
Taxon 3
Taxon 1
Taxon 2
3rd tree

Fig. 3. Impact of the branch lengths on the recovery ratio of the correct tree in the maximum-likelihood analysis with HKY + Γ model. I simulated 1,000 nucleotide-long sequence data with the difference of AT content across taxa of ≈20% and Ts/Tv ratio (κ) of 2.0 based on the 4-taxon model tree. 40 x 40 combinations of branch lengths of *a* and *b* of the model tree were examined. For each combination, I analyzed 100 replicates. The color of each cell in the matrix indicates the success rate.

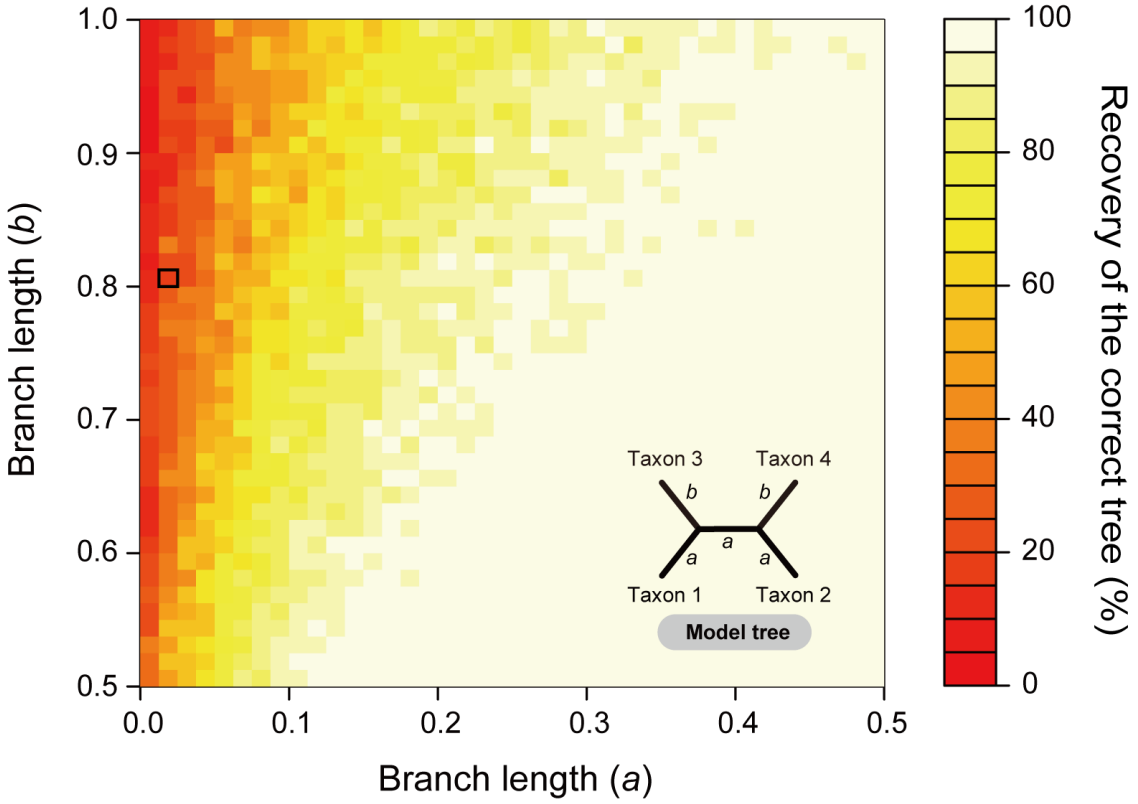

1.0
0.9
0.8
0.7
0.6
0.5
Branch length (b)
0.0
0.1
0.2
0.3
0.4
0.5
Branch length (a)
Taxon 3
Taxon 4
b
b
a
a
a
Taxon 1
Taxon 2
Model tree
100
80
60
40
20
0
Recovery of the correct tree (%)

Fig. 4. Impacts of the difference in AT content across taxa (ΔAT) and Ts/Tv ratio (κ) on the recovery rate of the correct tree. **(A)** Analysis of 4-taxon data simulated with κ = 2.0. I prepared 11 sets of 500 replicates of 1000 nt-long sequence data simulated with different ΔAT%. The simulated data were subjected to the ML analyses with the HKY + Γ model (HKY; black circles) and the GG98 + Γ model (GG98; green squares). I 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 the recoded data 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. I 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.

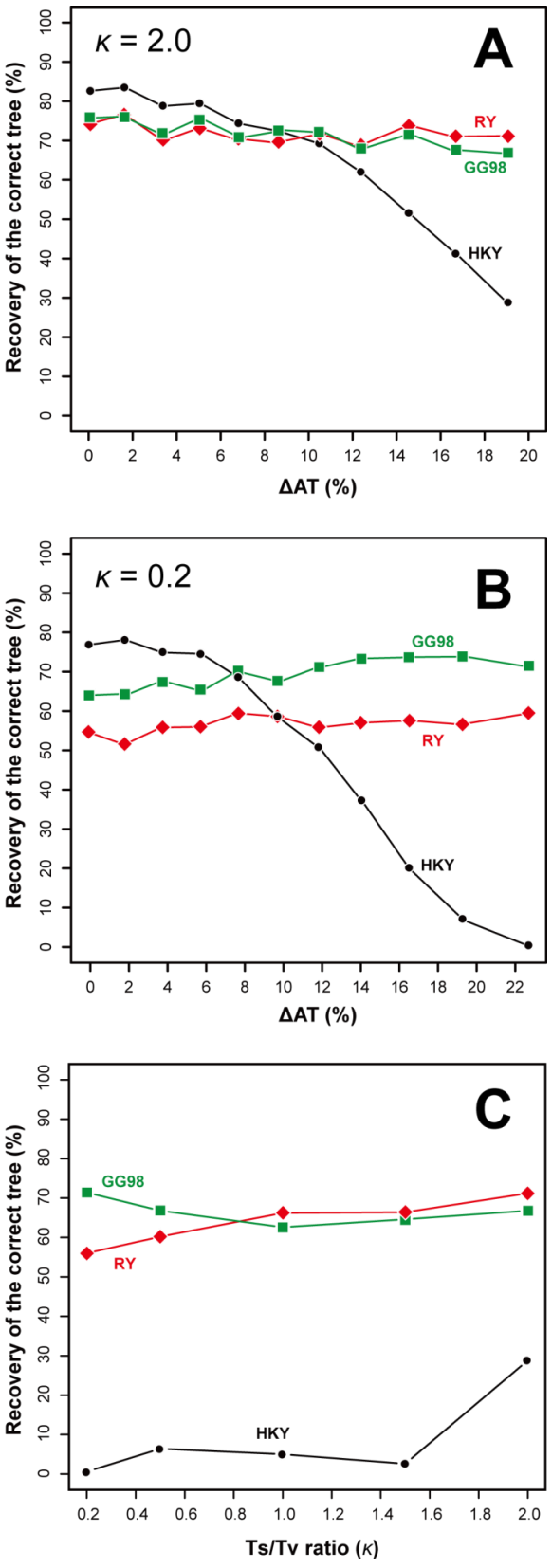

κ = 2.0
A
RY
GG98
HKY
Recovery of the correct tree (%)
ΔAT (%)
κ = 0.2
B
GG98
RY
HKY
Recovery of the correct tree (%)
ΔAT (%)
C
GG98
RY
HKY
Recovery of the correct tree (%)
Ts/Tv ratio (κ)

Fig. 5. Impact of the difference in AT content (ΔAT%) across tree on the recovery rate of 'LBA' tree, in which rapidly-evolving Taxa 3 and 4 group together (see Fig. 1B). The details of these figures are same as those in Fig. 2, except I plotted the recovery rates of LBA tree here.

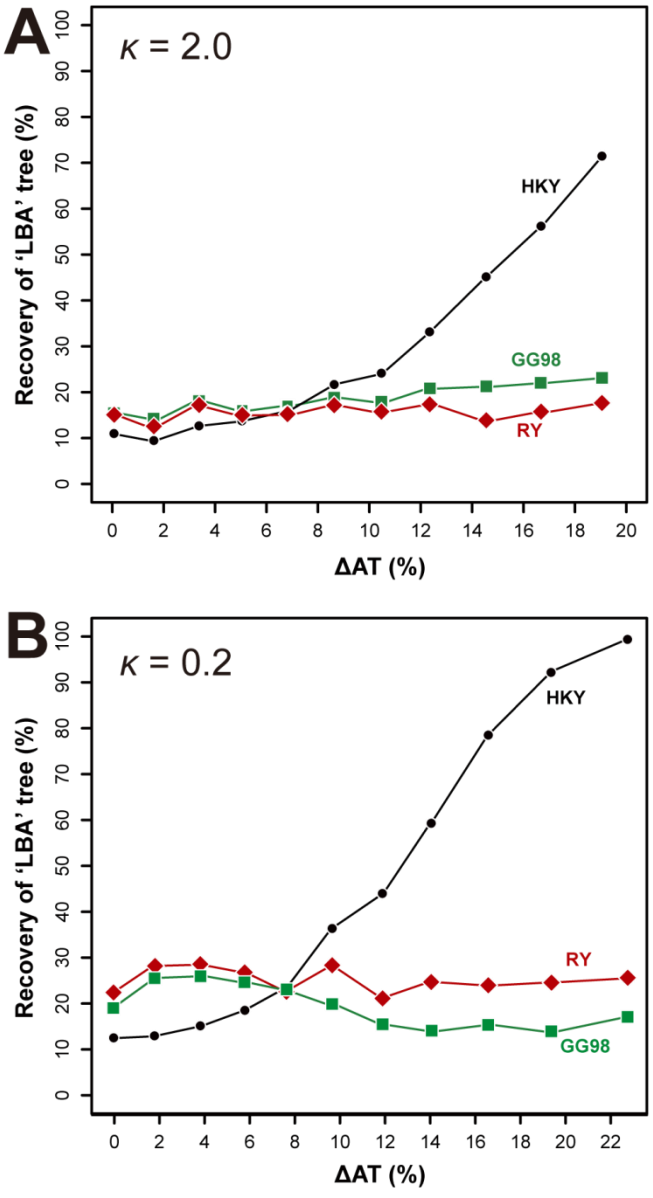

A
κ = 2.0
Recovery of 'LBA' tree (%)
HKY
GG98
RY
ΔAT (%)
B
κ = 0.2
Recovery of 'LBA' tree (%)
HKY
RY
GG98
ΔAT (%)

Fig. 6. Impact of data size on the recovery rate of the correct tree. **(A)** I 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 Ts/Tv ratio (κ) of 2.0. Five hundred replicates were simulated for each data point. The simulated data were subjected to the ML analyses with the HKY + Γ model (HKY; black circles) and the GG98 + Γ model (GG98; green squares). I 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 the recoded data 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 $\kappa = 0.2$.

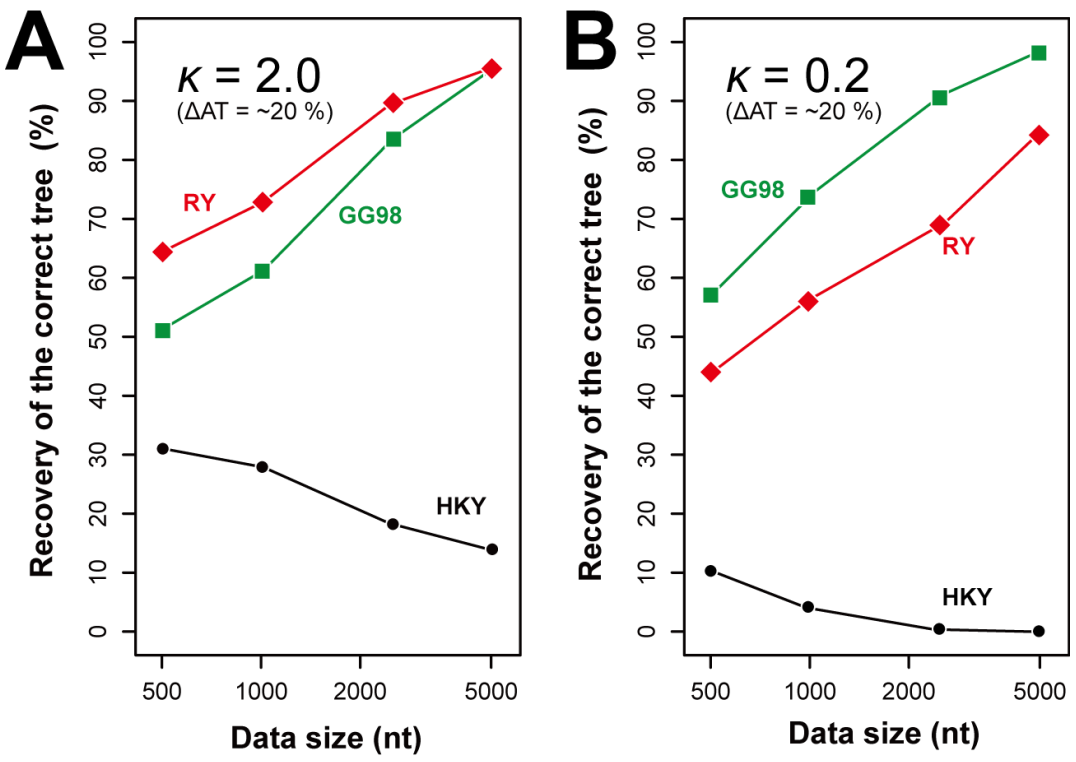

A
κ = 2.0
(ΔAT = ~20 %)
RY
GG98
HKY
Recovery of the correct tree (%)
0
10
20
30
40
50
60
70
80
90
100
500
1000
2000
5000
Data size (nt)
B
κ = 0.2
(ΔAT = ~20 %)
GG98
RY
HKY
Recovery of the correct tree (%)
0
10
20
30
40
50
60
70
80
90
100
500
1000
2000
5000
Data size (nt)

Fig. 7. Impact of complex base composition on the ML analyses with RY-coding and NH models. We simulated 1000 nt sequence data with AT content across a tree of ≈20% and Ts/Tv 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 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 non-homogeneous HKY + Γ model. The recovery of the correct and 'LBA' tree (see Fig. 1B) are shown as closed and open bars, respectively.

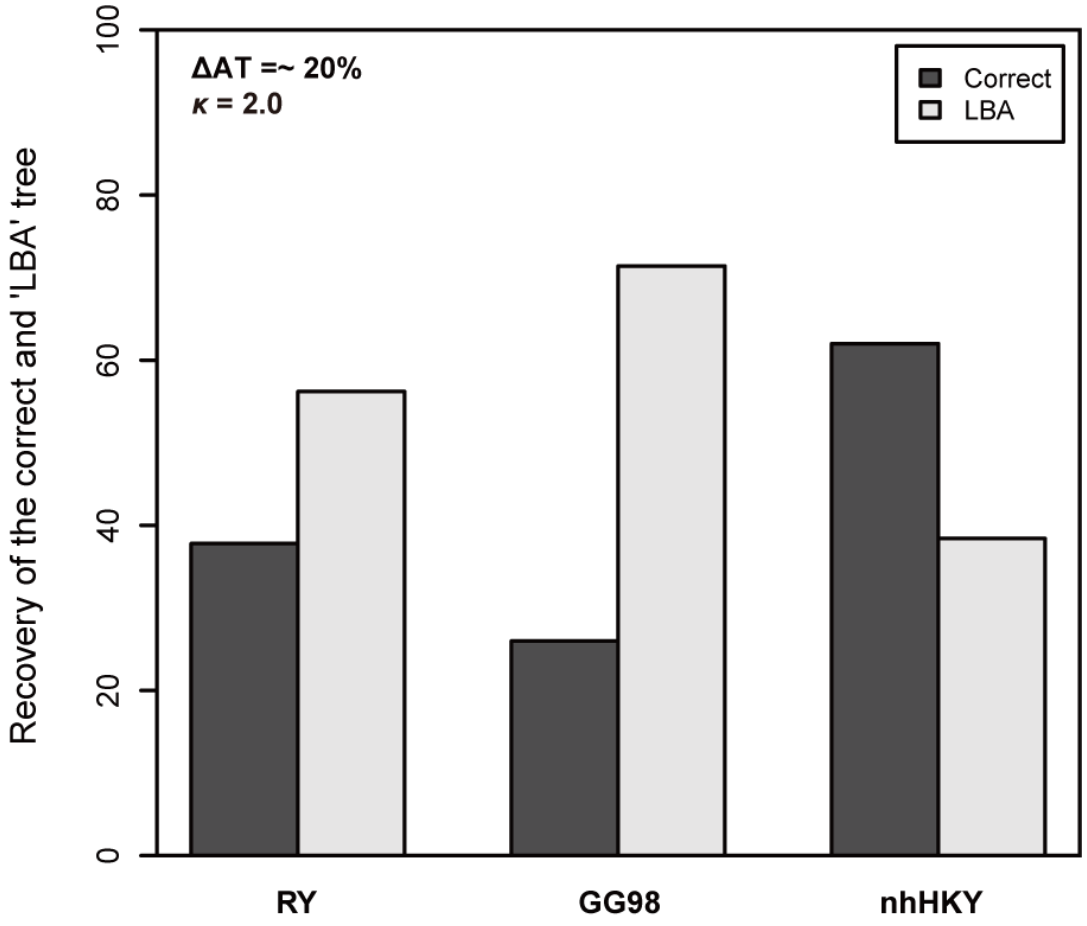

ΔAT =~ 20%
κ = 2.0
Correct
LBA
Recovery of the correct and 'LBA' tree
0
20
40
60
80
100
RY
GG98
nhHKY

Fig. 8. Impact of Ts/Tv ratio (κ) in the data simulation on the ML analyses of the RY-recoded data. **(A)** Difference in site pattern between the sequence data simulated with κ of 2.0 and those simulated with κ of 0.2 (shown as open and closed bars, respectively). I simulated a 50,000 nt-long simulated data, and recoded it 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. I optimized the branch lengths of the correct tree, in which rapidly evolving Taxa 3 and 4 are separated (see Fig. 2). 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.

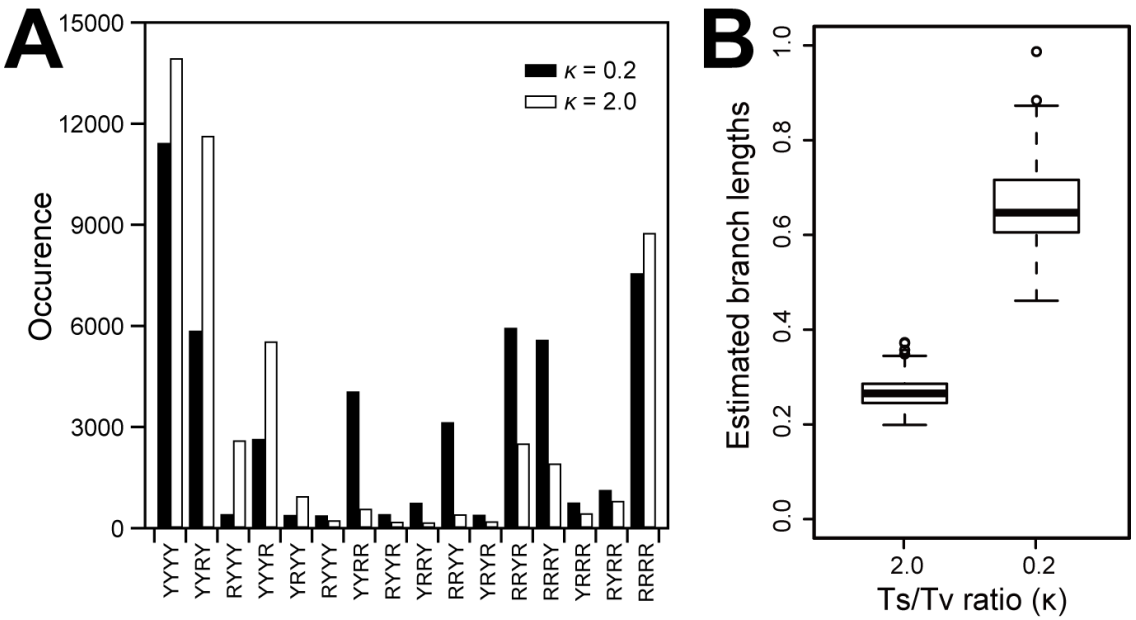
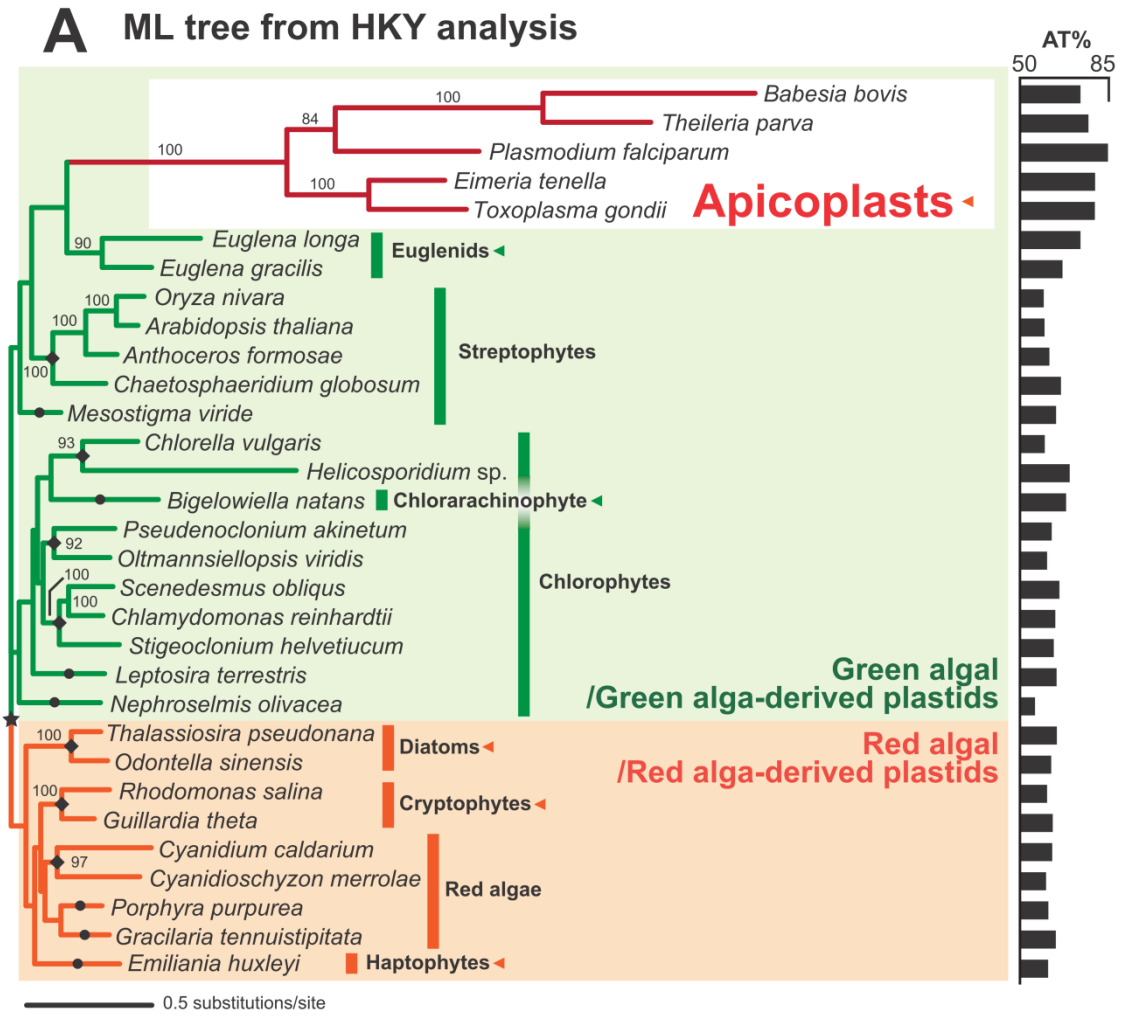

A
B
Occurence
15000
12000
9000
6000
3000
0
κ = 0.2
κ = 2.0
YYYY
YYRY
RYYY
YYYR
YRYY
RYYY
YYRR
RYYR
YRRY
RRYY
YRYR
RRYR
RRRY
YRRR
RYRR
RRRR
Estimated branch lengths
0.0
0.2
0.4
0.6
0.8
1.0
2.0
0.2
Ts/Tv ratio (κ)

Fig. 9. The ML tree from the 5-gene alignment with the homogeneous (HKY) model and two competing hypotheses for the origin of apicoplasts. **(A)** The ML 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 proportion larger than 50% is shown for each node. **(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. Abbreviations: N, host nucleus; n, endosymbiotic algal nucleus; P, plastid; Ap, apicoplast. Note that the nucleus of the endosymbiotic alga (n) has disappeared in modern apicomplexan cells.


A
ML tree from HKY analysis
AT%
50
85
Babesia bovis
Theileria parva
Plasmodium falciparum
Eimeria tenella
Toxoplasma gondii
Apicoplasts
Euglena longa
Euglena gracilis
Euglenids
Oryza nivara
Arabidopsis thaliana
Anthoceros formosae
Chaetosphaeridium globosum
Mesostigma viride
Streptophytes
Chlorella vulgaris
Helicosporidium sp.
Bigelowiella natans
Chlorarachinophyte
Pseudenoclonium akinetum
Oltmannsiellopsis viridis
Scenedesmus obliqus
Chlamydomonas reinhardtii
Stigeoclonium helvetiucum
Leptosira terrestris
Nephroselmis olivacea
Chlorophytes
Green algal
/Green alga-derived plastids
Thalassiosira pseudonana
Odontella sinensis
Diatoms
Rhodomonas salina
Guillardia theta
Cryptophytes
Cyanidium caldarium
Cyanidioschyzon merrolae
Porphyra purpurea
Gracilaria tennuistipitata
Red algae
Emiliania huxleyi
Haptophytes
Red algal
/Red alga-derived plastids
0.5 substitutions/site

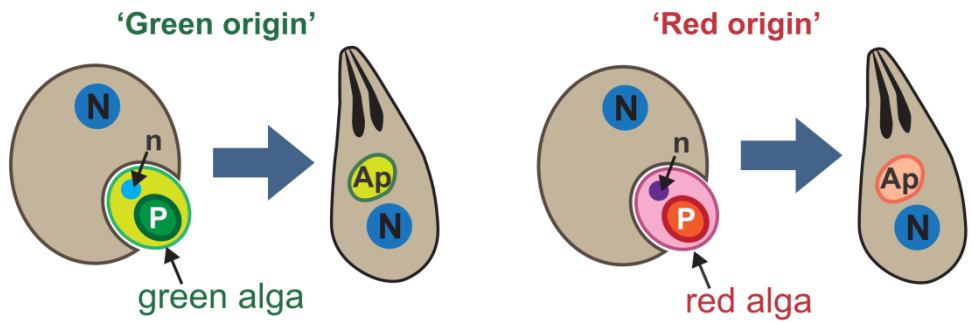
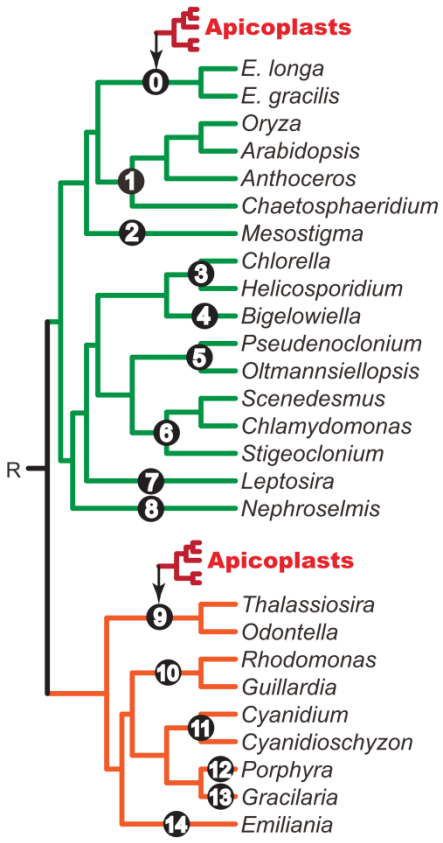
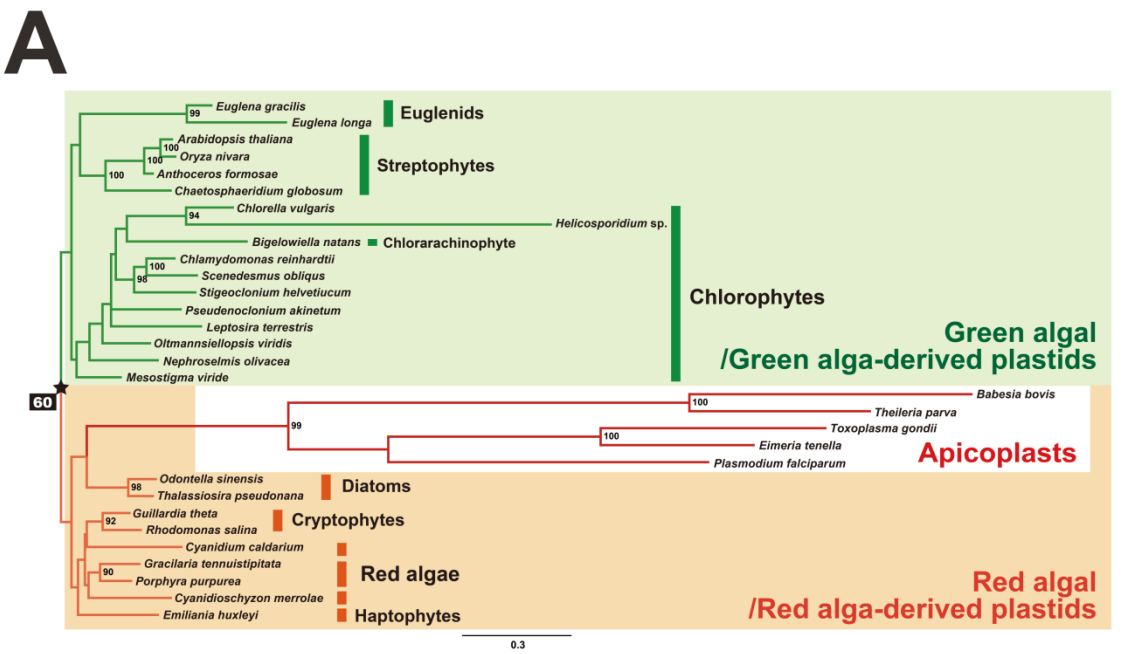

B
Two hypotheses for the origin of apicoplasts
'Green origin'
'Red origin'
N
n
P
Ap
N
green alga
red alga

Fig. 10. Tree ln*L* comparison. The phylogram (left) is created from the tree topology shown in Fig. 9A by pruning the entire apicoplast clade. The apicoplast clade was then re-grafted to positions labeled 0–14 to generate the trees assessed in this comparison. For instance, Trees 0 and 9 were generated by re-grafting 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 ln*L* calculation based on GG98 model is shown as 'R.' The table on the right provides the ln*L* value of the best tree among the 15 test trees, and the differences in ln*L* 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 CF + Γ model based on RY-recoded data are listed in the fourth column (labeled as RY), and ln*L* scores caluculated from the GG98 + Γ (non-homogeneous) model are listed in the fifth column (labeled as GG98).



	Apicoplast origin	HKY	RY	GG98
Tree 0	Green	BEST (-51941.54)	1.60	5.65
Tree 1	Green	13.06	5.40	9.42
Tree 2	Green	18.11	5.31	9.98
Tree 3	Green	30.71	13.07	27.39
Tree 4	Green	26.45	12.48	24.16
Tree 5	Green	35.98	11.44	26.86
Tree 6	Green	35.98	11.44	28.27
Tree 7	Green	30.50	11.42	21.78
Tree 8	Green	26.68	7.36	16.66
Tree 9	Red	7.47	BEST (-26573.15)	BEST (-50667.26)
Tree 10	Red	13.92	6.04	1.44
Tree 11	Red	14.12	6.04	2.03
Tree 12	Red	22.40	9.73	10.96
Tree 13	Red	21.04	9.78	9.91
Tree 14	Red	11.85	3.73	0.70

Fig. 11. Results from the MLBP analyses incorporating RY-coding and GG98 model. **(A)** The ML tree inferred from the CF + Γ model with RY-recoded data using PhyML. **(B)** The ML tree inferred from the GG98 + Γ model with nt sequence data using the *shake_nh* program in NHML. Details are same as described in Fig. 9A.


A
Euglena gracilis
Euglena longa
Euglenids
Arabidopsis thaliana
Oryza nivara
Anthoceros formosae
Chaetosphaeridium globosum
Streptophytes
Chlorella vulgaris
Helicosporidium sp.
Bigelowiella natans
Chlorarachinophyte
Chlamydomonas reinhardtii
Scenedesmus obliqus
Stigeoclonium helvetiucum
Pseudenoclonium akinetum
Leptosira terrestris
Oltmannsiellopsis viridis
Nephroselmis olivacea
Mesostigma viride
Chlorophytes
Green algal
/Green alga-derived plastids
60
Babesia bovis
Theileria parva
Toxoplasma gondii
Eimeria tenella
Plasmodium falciparum
Apicoplasts
Odontella sinensis
Thalassiosira pseudonana
Diatoms
Guillardia theta
Rhodomonas salina
Cryptophytes
Cyanidium caldarium
Gracilaria tennuistipitata
Porphyra purpurea
Cyanidioschyzon merrolae
Red algae
Emiliania huxleyi
Haptophytes
Red algal
/Red alga-derived plastids
0.3

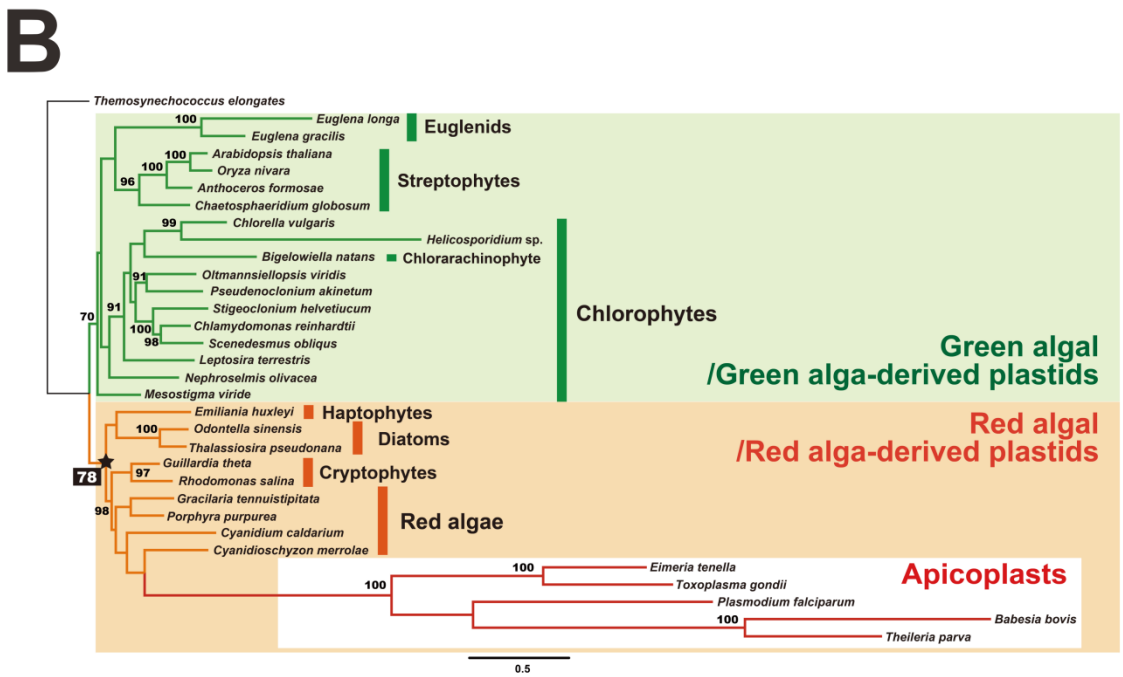

B
Themosynechococcus elongates
Euglena longa
Euglena gracilis
Euglenids
Arabidopsis thaliana
Oryza nivara
Anthoceros formosae
Chaetosphaeridium globosum
Streptophytes
Chlorella vulgaris
Helicosporidium sp.
Bigelowiella natans
Chlorarachinophyte
Oltmannsiellopsis viridis
Pseudenoclonium akinetum
Stigeoclonium helvetiucum
Chlamydomonas reinhardtii
Scenedesmus obliqus
Leptosira terrestris
Nephroselmis olivacea
Mesostigma viride
Chlorophytes
Green algal
/Green alga-derived plastids
Emiliania huxleyi
Haptophytes
Odontella sinensis
Thalassiosira pseudonana
Diatoms
Guillardia theta
Rhodomonas salina
Cryptophytes
Gracilaria tennuistipitata
Porphyra purpurea
Cyanidium caldarium
Cyanidioschyzon merrolae
Red algae
Red algal
/Red alga-derived plastids
78
Eimeria tenella
Toxoplasma gondii
Plasmodium falciparum
Babesia bovis
Theileria parva
Apicoplasts
0.5

Fig. 12. Newton-Raphson algorithm for the calculation of the maximum ln*L* score for a given tree. Pseudo-code indicates each step (i~viii) to calculate the maximum-likelihood estimates of branch lengths and model parameters. The bold- and italic-sentences represent control statements in the program; *WHILE* means the loop in which processes written between *WHILE* and *ENDWHILE* are repeated as long as the test condition is true; *FOR* means the loop in which processes written between *FOR* and *ENDFOR* are repeated in a range of specified numbers; *RETURN* means that the algorithm finishes outputting the result.

Newton-Raphson algorithm

calculate initial log-likelihood ($\ln L$) of a tree from randomly determined values of branch lengths and model parameters – (i)

WHILE (first step or $\Delta \ln L \geq \varepsilon$) – (ii)

FOR (number of parameters to be estimated) – (iii)

FOR (number of positions) – (iv)

calculate 1st and 2nd derivatives of site-ln$L$s with respect to a parameter (θ) to be optimized.

ENDFOR

ENDFOR

update parameters – (v)

calculate current ln$L$ from updated parameters – (vi)

calculate the difference between current ln$L$ score and previous one (= $\Delta \ln L$) – (vii)

ENDWHILE

RETURN current ln$L$ as maximum ln$L$ – (viii)

Fig. 13. The parallel computation scheme for multiple trees. The ln*L* scores for *N* alternative trees are calculated in parallel by partitioning *MPI_COMM_WORLD* into several sub-communicators, which control partial group of MPI processes and OpenMP threads respectively. Each sub-communicator is then assigned to compute the ln*L* score for different group of trees.

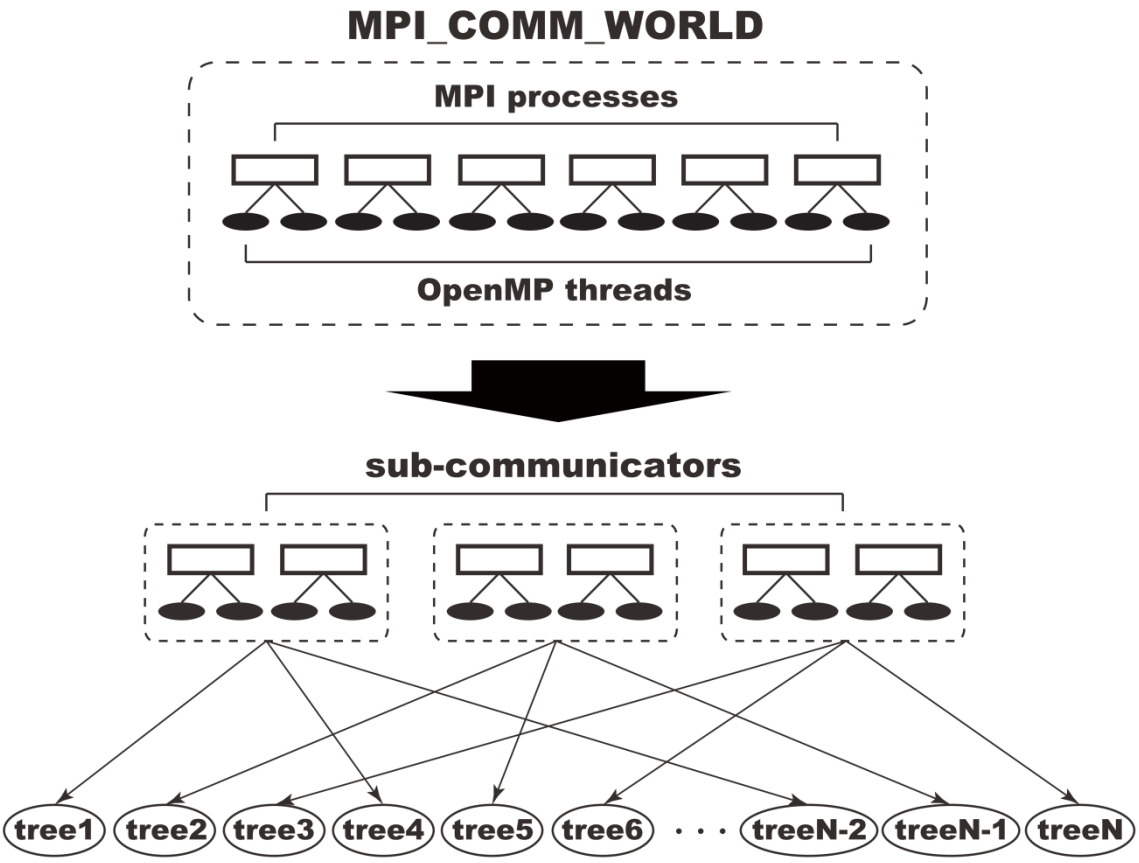

MPI_COMM_WORLD
MPI processes
OpenMP threads
sub-communicators
tree1
tree2
tree3
tree4
tree5
tree6
treeN-2
treeN-1
treeN

Fig. 14. The 66-taxon trees for benchmark analyses. **(A)**. The model tree for sequence generation. The tree is shown as phylogram but lengths for black-colored branches are actually set to 0.05, while those for red-colored branches are set to 1.0. 'R' means the root position of the tree. Red arrowheads indicate the point to change the AT content from 50% to 90% during sequence simulation. Thus, sequences evolve following extremely high AT content and rapid substitution rates on the corresponding branches. Squares represent branches which would be bisected to generate the 130-taxon tree for (see IV-2-4). **(B)**. The ML tree inferred from the homogeneous GTR + Γ model. This tree is then modified to generate 24 alternative trees by re-grafting the clade of taxa 63 and 64 to alternative positions (highlighted by stars). Details are described in IV-2-4.

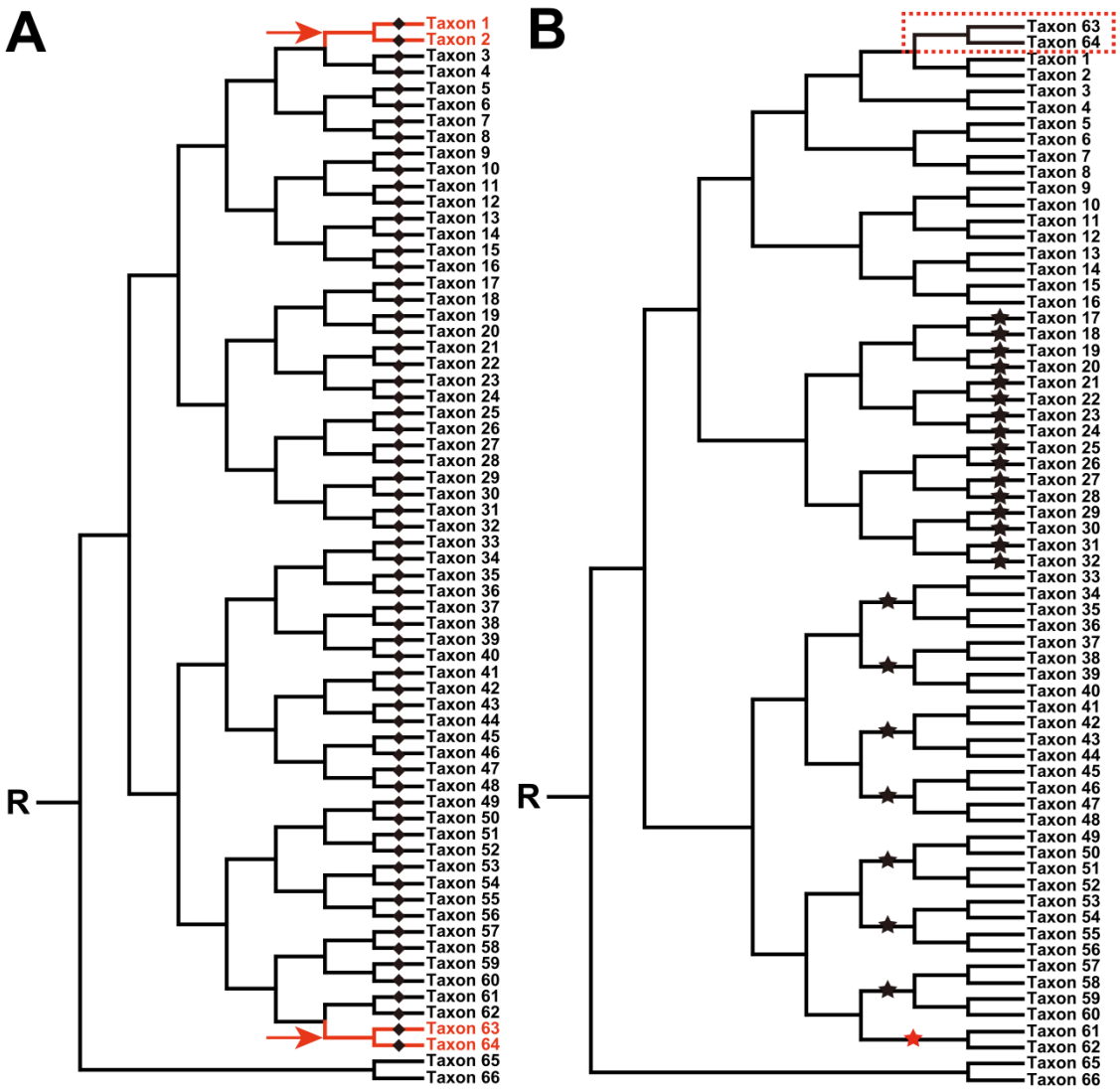

A
R
Taxon 1
Taxon 2
Taxon 3
Taxon 4
Taxon 5
Taxon 6
Taxon 7
Taxon 8
Taxon 9
Taxon 10
Taxon 11
Taxon 12
Taxon 13
Taxon 14
Taxon 15
Taxon 16
Taxon 17
Taxon 18
Taxon 19
Taxon 20
Taxon 21
Taxon 22
Taxon 23
Taxon 24
Taxon 25
Taxon 26
Taxon 27
Taxon 28
Taxon 29
Taxon 30
Taxon 31
Taxon 32
Taxon 33
Taxon 34
Taxon 35
Taxon 36
Taxon 37
Taxon 38
Taxon 39
Taxon 40
Taxon 41
Taxon 42
Taxon 43
Taxon 44
Taxon 45
Taxon 46
Taxon 47
Taxon 48
Taxon 49
Taxon 50
Taxon 51
Taxon 52
Taxon 53
Taxon 54
Taxon 55
Taxon 56
Taxon 57
Taxon 58
Taxon 59
Taxon 60
Taxon 61
Taxon 62
Taxon 63
Taxon 64
Taxon 65
Taxon 66
B
R
Taxon 63
Taxon 64
Taxon 1
Taxon 2
Taxon 3
Taxon 4
Taxon 5
Taxon 6
Taxon 7
Taxon 8
Taxon 9
Taxon 10
Taxon 11
Taxon 12
Taxon 13
Taxon 14
Taxon 15
Taxon 16
Taxon 17
Taxon 18
Taxon 19
Taxon 20
Taxon 21
Taxon 22
Taxon 23
Taxon 24
Taxon 25
Taxon 26
Taxon 27
Taxon 28
Taxon 29
Taxon 30
Taxon 31
Taxon 32
Taxon 33
Taxon 34
Taxon 35
Taxon 36
Taxon 37
Taxon 38
Taxon 39
Taxon 40
Taxon 41
Taxon 42
Taxon 43
Taxon 44
Taxon 45
Taxon 46
Taxon 47
Taxon 48
Taxon 49
Taxon 50
Taxon 51
Taxon 52
Taxon 53
Taxon 54
Taxon 55
Taxon 56
Taxon 57
Taxon 58
Taxon 59
Taxon 60
Taxon 61
Taxon 62
Taxon 65
Taxon 66

Fig. 15. Changes of the total execution time against the number of CPU cores in the analyses of **(A)** small and **(B)** large 66-taxon datasets, as well as in the analysis of **(C)** 130-taxon dataset. Horizontal axes mean total execution time (s) for computing all trees based on 66-taxon and 130-taxon datasets, and vertical axes mean the number of CPU cores used in each run.

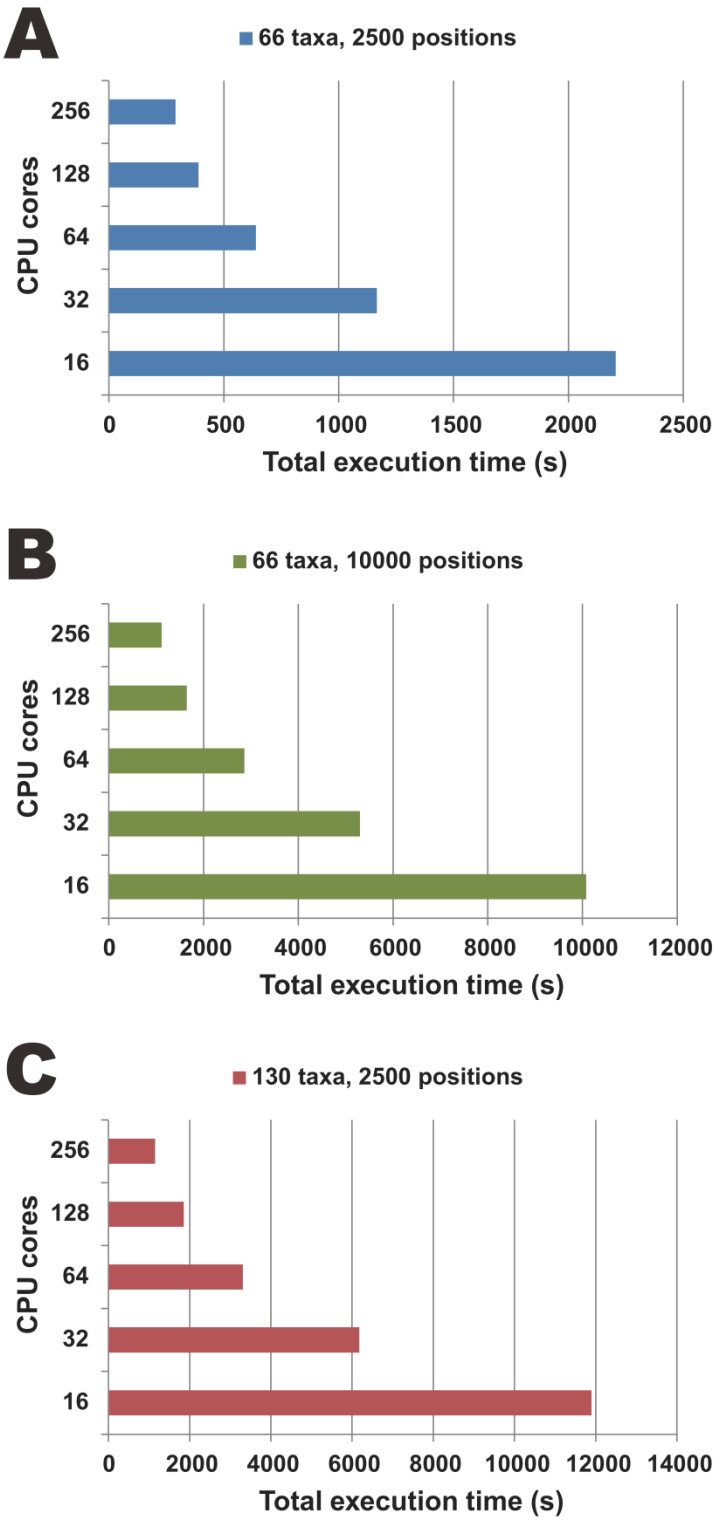

A
66 taxa, 2500 positions
CPU cores
256
128
64
32
16
0
500
1000
1500
2000
2500
Total execution time (s)
B
66 taxa, 10000 positions
CPU cores
256
128
64
32
16
0
2000
4000
6000
8000
10000
12000
Total execution time (s)
C
130 taxa, 2500 positions
CPU cores
256
128
64
32
16
0
2000
4000
6000
8000
10000
12000
14000
Total execution time (s)

Fig. 16. Plots of speeding-up ratios against the numbers of CPU cores. Speeding-up ratios were measured up to on 256 CPU cores normalizing them by the run time on 16 CPU cores. Each ratio is plotted for small and large 66-taxon datasets (blue circles and green diamonds respectively), as well as 130-taxon dataset (red squares). Black line represents ideal, linear speeding-up with parallel efficiency equal to 1.0.

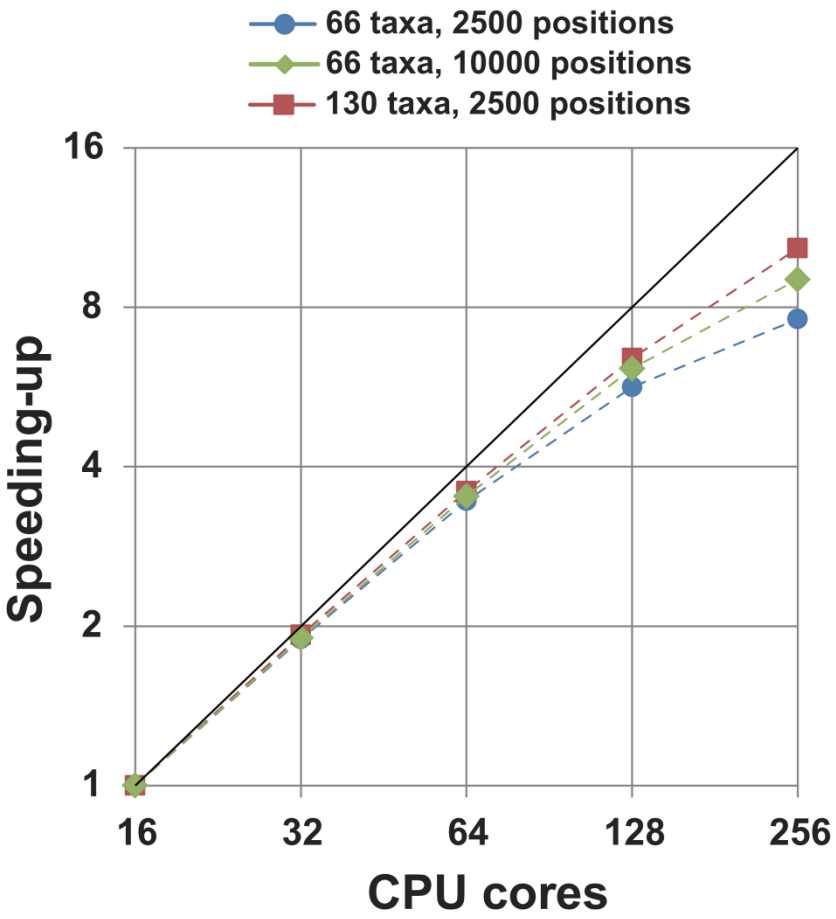

66 taxa, 2500 positions
66 taxa, 10000 positions
130 taxa, 2500 positions
Speeding-up
16
8
4
2
1
16
32
64
128
256
CPU cores

Fig. 17. Breakdowns of the time for *MPI_Allgatherv* communication (henceforth designated as 'Comm time') and the substantial time for the ln*L* calculation (henceforth designated as 'CPU time'). Comm time and CPU time against the number of CPU cores were measured in the analyses of **(A)** small and **(B)** large 66-taxon datasets, and **(C)** 130-taxon dataset.

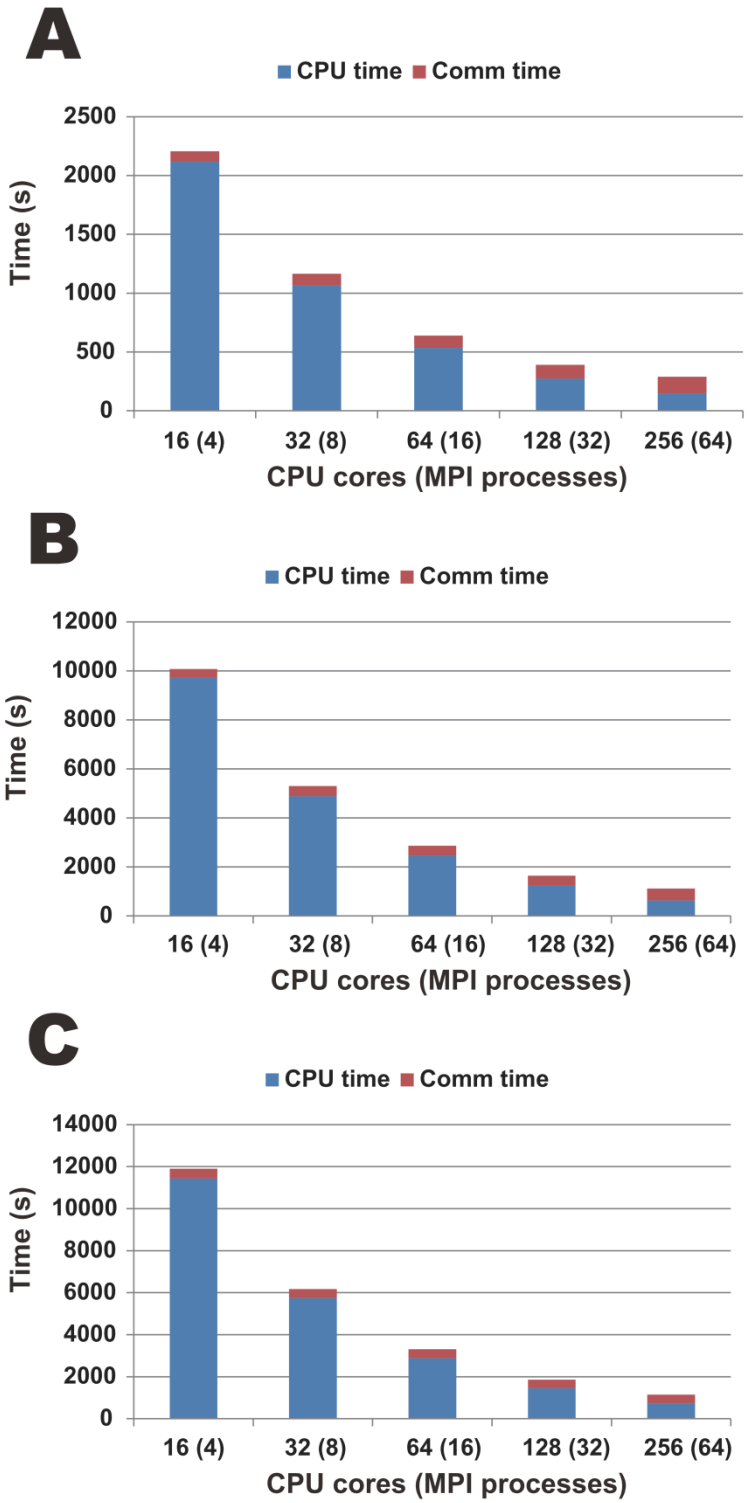
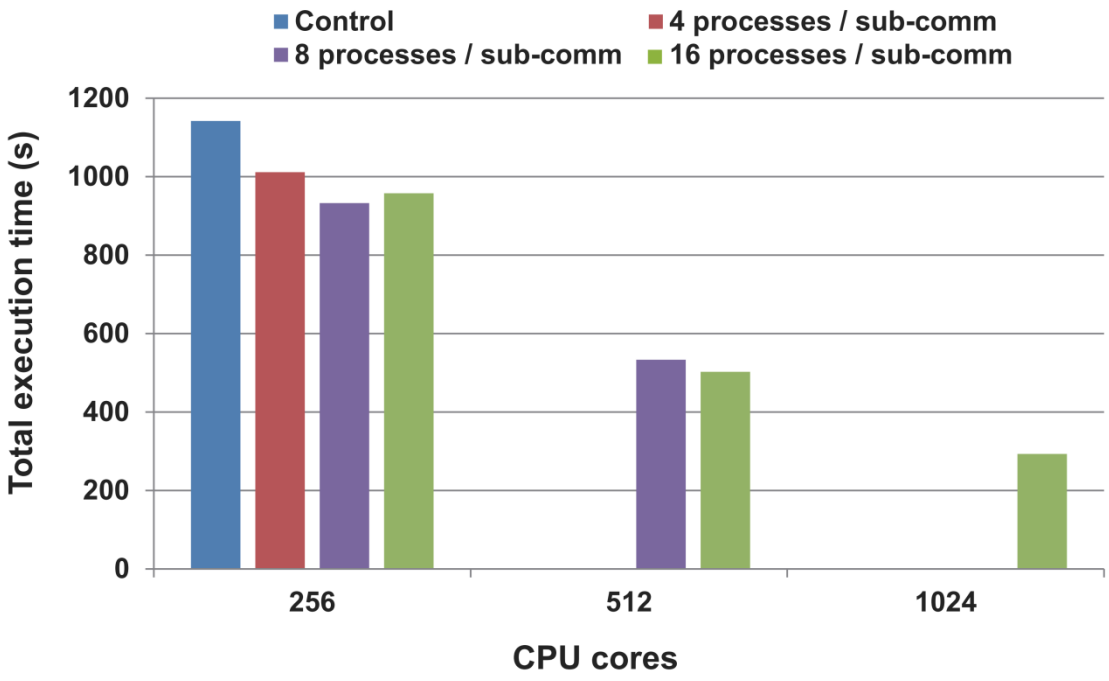

A
CPU time
Comm time
2500
2000
1500
1000
500
0
Time (s)
16 (4)
32 (8)
64 (16)
128 (32)
256 (64)
CPU cores (MPI processes)
B
CPU time
Comm time
12000
10000
8000
6000
4000
2000
0
Time (s)
16 (4)
32 (8)
64 (16)
128 (32)
256 (64)
CPU cores (MPI processes)
C
CPU time
Comm time
14000
12000
10000
8000
6000
4000
2000
0
Time (s)
16 (4)
32 (8)
64 (16)
128 (32)
256 (64)
CPU cores (MPI processes)

Fig. 18. Further speeding-up by the implementation of the parallel computation of multiple trees for the 130-taxon dataset. I applied three different partition schemes for *MPI_COMM_WORLD*. For each partition scheme, total execution time for computing 48 alternative trees is shown as red, purple, and green bars. The total execution time for the control run, where all MPI processes were assigned to compute the same tree without partitioning *MPI_COMM_WORLD*, is shown as blue bar.


Control
4 processes / sub-comm
8 processes / sub-comm
16 processes / sub-comm
Total execution time (s)
1200
1000
800
600
400
200
0
256
512
1024
CPU cores