mPartition: A Model-Based Method for Partitioning Alignments

Thu Le Kim & Vinh Le Sy

Journal of Molecular Evolution

ISSN 0022-2844

J Mol Evol
DOI 10.1007/s00239-020-09963-z

Springer

Journal of Molecular Evolution
https://doi.org/10.1007/s00239-020-09963-z

ORIGINAL ARTICLE

mPartition: A Model-Based Method for Partitioning Alignments

Thu Le Kim[1,2] · **Vinh Le Sy**[1]

Received: 31 December 2019 / Accepted: 8 August 2020

Abstract
Maximum likelihood (ML) analysis of nucleotide or amino-acid alignments is widely used to infer evolutionary relationships among species. Computing the likelihood of a phylogenetic tree from such alignments is a complicated task because the evolutionary processes typically vary across sites. A number of studies have shown that partitioning alignments into sub-alignments of sites, where each sub-alignment is analyzed using a different model of evolution (e.g., GTR + I + G), is a sensible strategy. Current partitioning methods group sites into subsets based on the inferred rates of evolution at the sites. However, these do not provide sufficient information to adequately reflect the substitution processes of characters at the sites. Moreover, the site rate-based methods group all invariant sites into one subset, potentially resulting in wrong phylogenetic trees. In this study, we propose a partitioning method, called mPartition, that combines not only the evolutionary rates but also substitution models at sites to partition alignments. Analyses of different partitioning methods on both real and simulated datasets showed that mPartition was better than the other partitioning methods tested. Notably, mPartition overcame the pitfall of grouping all invariant sites into one subset. Using mPartition may lead to increased accuracy of ML-based phylogenetic inference, especially for multiple loci or whole genome datasets.

Keywords Alignment partitioning · Maximum likelihood phylogenetic inference · Substitution model · Site rate model

Background

Phylogenetic inference is a powerful approach to study the evolutionary relationships among species. The maximum likelihood (ML) method is among the most popular approaches to infer phylogenetic trees from nucleotide and amino-acid sequences (Felsenstein 2003; Lemey et al. 2009). The accuracy of ML-based phylogenetic inference relies on a number of factors including the size of alignments (i.e., the number of sites and sequences), tree building methods (e.g., IQ-TREE or PhyML), and models of sequence evolution (e.g., GTR + I + G4 or HKY + G4). The advancement of sequencing technologies has created large datasets for inferring phylogenetic trees. Efficient ML methods have been developed to build phylogenetic trees from large datasets; these include PhyML (Guindon and Gascuel 2003), IQPNNI (Vinh and von Haeseler 2004), RAxML (Stamatakis 2015), and IQ-TREE (Minh et al. 2020). Using different models of evolution to analyze a given dataset might produce significantly different trees (Frandsen et al. 2015; Kainer and Lanfear 2015; Le and Gascuel 2008; Rota et al. 2018). The misspecification of evolutionary models often results in systematic errors, such as strong supports of incorrect clades in the constructed trees (Hoang et al. 2017), especially for large datasets, including multiple loci or whole genomes (Kumar et al. 2012; Rodréguez-Ezpeleta et al. 2007).

It is well known that the evolutionary processes among sites rarely are homogeneous (i.e., the evolutionary rates often vary among sites). Currently, two main approaches to model rate heterogeneity among sites are the mixture model approach (Le et al. 2012; Pagel and Meade 2004) and the partitioning approach (Frandsen et al. 2015; Lanfear et al. 2012; Nylander et al. 2004; Rota et al. 2018). The mixture model approach uses several models to calculate the likelihood value of each site. On the other hand, the partitioning approach accounts for the heterogeneity among sites by classifying sites into several disjoint subsets (a partition scheme)

Handling Editor: **Arndt von Haeseler**.

✉ Vinh Le Sy
vinhls@vnu.edu.vn

[1] University of Engineering and Technology, Vietnam National University Hanoi, 144 Xuan Thuy, Cau Giay, Hanoi 10000, Vietnam

[2] Hanoi University of Science and Technology, 1st Dai Co Viet, Hai Ba Trung, Hanoi 10000, Vietnam

Published online: 31 August 2020

such that sites in the same subset are assumed to evolve under the same model of evolution. In other words, sites that are under the same evolutionary process are grouped together. The partitioning approach has been more widely used than the mixture model approach in ML phylogenetic analyses due to its computational efficiency and software availability. However, the recent implementation of a mixture model, i.e., GHOST (Crotty et al. 2020), in the IQ-TREE package allows users to efficiently build ML trees with the mixture model.

Simple partitioning methods use biological properties of sites such as gene boundaries or codon positions (Brandley et al. 2005; Lartillot and Philippe 2004; Nylander et al. 2004; Yang 1996) to group sites into similar subsets. The codon-based method groups sites at the same codon position into one subset. Although sites at the same codon position share some common features, the assumption that all sites at the same codon position evolved under the same model of evolution is not always biologically realistic. They might evolve at different rates and follow different substitution patterns. Moreover, the biological information is not always available for partitioning sites, so computational methods are required to automatically partition alignments.

A number of studies have proposed computational methods to automatically cluster sites based on the evolutionary rates of sites (Frandsen et al. 2015; Rota et al. 2018). The site rate-based methods cluster sites into groups based on the assumption that sites have similar evolutionary rates should be in the same subset. Specifically, the *k*-means algorithm (Frandsen et al. 2015) iteratively clusters sites into subsets of similar site rates. Grouping all invariant sites into one subset is a critical pitfall of the *k*-means algorithm because sites might be invariant in the dataset under the study, but might vary in larger datasets. The pitfall might increase the likelihood of tree, but result in biased trees (Baca et al. 2017; Rota et al. 2018). Therefore, since 2017 the *k*-means algorithm is no longer recommended. To diminish the problem, the RatePartition algorithm (Rota et al. 2018) employs a simple formula to divide sites into subsets such that the first subset includes not only all invariant sites but also some additional sites with the slowest evolutionary rates. During a preliminary study of this algorithm, we found that the accuracy of RatePartition was affected by the division factor parameter d, which controls the number of partitions in the partition scheme. A higher d value results in a greater number of partitions, hence, the change of d value affects the accuracy of RatePartition.

The evolutionary rates of different sites provide useful signals for partitioning sites; however, that information is unable to represent the substitution processes of characters (nucleotide or amino-acids). The substitution process at a site can be modeled by a Markov process and described by a substitution model M, which represents the instantaneous substitution rates between characters. The substitution models at sites provide crucial signals for partitioning sites. Our proposed mPartition method combines information from both the evolutionary rates at sites (i.e., the inferred rate of evolution at the sites) and substitution models (i.e., the substitution rates among nucleotides or amino-acids) to cluster sites into subsets such that sites in the same subset have a similar evolutionary rate and substitution model. We examined the accuracy of mPartition and other partitioning methods on both nucleotide datasets obtained from the paper describing the RatePartition method (Rota et al. 2018) and amino-acid datasets consisting of multiple loci alignments selected from previous studies. We analyzed the distribution of invariant sites from the mPartition method to assess its ability to overcome the pitfall of grouping all invariants into one subset.

Methods and Material

Methods

Given a multiple sequence alignment $A = \{a_1, \ldots, a_l\}$ of l sites, the ML tree reconstruction method normally determines a tree T and a substitution model M and a site rate model V to maximize the likelihood $L(A|T, M, V)$. The substitution model M represents substitution rates between characters during the evolution that is usually simplified and modeled by a time-homogeneous and time-reversible Markov process. Technically, the substitution model M is described by a $n \times n$ instantaneous substitution rate matrix $Q = \{q_{xy}\}$, where q_{xy} represents the number of substitutions between two different characters x and y per time unit ($n = 4$ for nucleotides; and $n = 20$ for amino-acids). The site rate model V describes the rate heterogeneity among sites (i.e., different evolutionary rates among sites). Typically, a site rate model V combines a Γ distribution rate model and an invariant rate model (Yang 1993). Alternatively, the probability-distribution-free model may be used to describe the rate heterogeneity among sites (Kalyaanamoorthy et al. 2017).

We assume that the substitution processes among sites are independent, so the likelihood value $L(A|T, M, V)$ can be calculated as follows:

$$L(A|T, M, V) = \prod_{i=1}^{l} L(a_i|T, M, V) \propto \prod_{i=1}^{l} P(a_i|T, M, V)$$

where the likelihood value $L(a_i|T, M, V)$ can be calculated by the conditional probability $P(a_i|T, M, V)$ of data a_i given tree T, substitution model M, and site rate model V.

The alignment partitioning methods divide sites into several disjoint subsets such that sites in the same subset are assumed to evolve under a similar site rate and substitution

model. The disjoint subsets are called a partition scheme. Let $\boldsymbol{S} = \{S_1, \ldots, S_p\}$ be a partition scheme of p subsets satisfying that each site of alignment A belongs to one and only one subset of partition scheme $\boldsymbol{S}$. Let $\boldsymbol{M} = \{M_1, \ldots, M_p\}$ be a set of substitution models, where M_i is the substitution model for subset S_i. Similarly, let $\boldsymbol{V} = \{V_1, \ldots, V_p\}$ be a set of site rate models, where V_i is the site rate model for subset S_i. The likelihood value $L(\boldsymbol{S}|T, \boldsymbol{M}, \boldsymbol{V})$ can be calculated as follows:

$$L(\boldsymbol{S}|T, \boldsymbol{M}, \boldsymbol{V}) = \prod_{i=1}^{p} \prod_{j=1}^{l_i} L\left(S_{ij}|T, M_i, V_i\right) \propto \prod_{i=1}^{p} \prod_{j=1}^{l_i} P\left(S_{ij}|T, M_i, V_i\right)$$

where l_i is the number of sites in subset S_i;S_{ij} is the data at the jth site of subset S_i; and $P(S_{ij}|T, M_i, V_i)$ is the conditional probability of data S_{ij}, given tree T, substitution model M_i, and site rate model V_i.

Given an alignment A, the partitioning methods determine the best partition scheme $\boldsymbol{S}$ for A by maximizing the likelihood value $L(\boldsymbol{S}|T, \boldsymbol{M}, \boldsymbol{V})$. Note that information-theoretic metrics [i.e., the corrected Akaike information criterion (AICc) (Hurvich and Tsai 1989) and the Bayesian information criterion (BIC) (Dziak et al. 2020; Schwarz 1978)] are normally used to measure the fitness of models $\boldsymbol{M}$ and $\boldsymbol{V}$ with different number of free parameters.

The mPartition algorithm starts from a partition scheme with only one subset consisting of all sites, and iteratively partitions current subsets into new smaller subsets to decrease the total BIC score of the current partition scheme. To partition a subset, the mPartition algorithm clusters its sites into three subsets (i.e., assigning sites with slow, medium, and fast evolutionary rates to slow-rate, medium-rate, and fast-rate subsets, respectively). This step is similar to site rate-based partitioning methods. The Tree Independent Generation of Evolutionary rates (TIGER) method (Cummins and McInerney 2011) is widely used to estimate the evolutionary rates of sites because it uses the composition of character patterns in the alignment without employing any tree to avoid the tree bias (Frandsen et al. 2015; Rota et al. 2018).

The key difference between mPartition and other site rate-based algorithms is the re-partitioning step based on the best-fit models (i.e., including both site rate model and substitution model) of the sites. Moreover, we apply the likelihood-mapping idea (Strimmer and von Haeseler 1997) to distribute invariant sites into different subsets proportional to their likelihood values in order to overcome the pitfall of grouping all invariant sites into one subset.

The mPartition algorithm is composed of four steps and described as follows:

1. **Initial step:** Let $\boldsymbol{S}$ be a partition scheme. Each subset in $\boldsymbol{S}$ is labeled either "partitioned" (i.e., not under the partitioning process anymore) or "partitioning" (i.e., being under the partitioning process). Initially, $\boldsymbol{S} = \{A\}$ (i.e., the initial scheme has only one subset containing all sites of alignment A), and the subset A is labeled "partitioning." Compute evolutionary rates for all sites of A using the TIGER algorithm. Construct a tree T from the alignment A.
2. **Partitioning by site rates:** Let S be a "partitioning" subset of $\boldsymbol{S}$. Let $r(S_j)$ be the rate of site S_j in S; r_{max} and r_{min} be the highest and lowest site rates for S, respectively. Note that the site with the fastest evolutionary rate has the smallest TIGER rate value. We cluster the sites of S into three subsets: high-rate subset (P_1), medium-rate subset (P_2), and low-rate subset (P_3). Technically, the site S_j is clustered into subset $g(S_j)$ as follows:

$$g\left(S_j\right) = \begin{cases} P_1 : r\left(S_j\right) < r_{min} + k \\ P_2 : r_{min} + k \le r\left(S_j\right) < r_{min} + 2 \times k \\ P_3 : r_{min} + 2 \times k \le r\left(S_j\right) \end{cases}$$

   where $k = \frac{r_{max} - r_{min}}{3}$ is one third the difference between r_{max} and r_{min}.
3. **Re-partitioning by models:** Reoptimize branch lengths of the tree T with respect to "partitioning" subset S. Determine the best-fit models for new partitioned subsets P_1, P_2, and P_3 using the ModelFinder algorithm (Kalyaanamoorthy et al. 2017). Let (M_1, V_1), (M_2, V_2), and (M_3, V_3) be the best-fit models for subsets P_1, P_2, and P_3, respectively. For each site S_j, we calculate likelihood values $L(S_j|T, M_1, V_1)$, $L(S_j|T, M_2, V_2)$ and $L(S_j|T, M_3, V_3)$. If S_j is a variant site, re-assign S_j to the highest likelihood subset. Otherwise (i.e., S_j contains only one nucleotide/amino-acid type), re-assign S_j to subset P_x with a probability p_x proportional to its likelihood value computed as follows:

$$p_x = L(S_j|T, M_x, V_x) / \sum_{v=1\ldots3} L(S_j|T, M_v, V_v)$$

   The re-assigning strategy will partition invariant sites into different subsets to overcome the pitfall of assigning all invariant sites into one subset. To avoid creating small subsets, if a subset P_x has less than 50 sites [the same threshold as used in (Tagliacollo and Lanfear 2018)], the subset P_x will be removed from the partition scheme by re-assigning all sites in P_x to other subsets.

   If the total BIC score of new subsets P_1, P_2, and P_3 is better than the BIC score of subset S, replace subset S by the new subsets P_1, P_2, and P_3 and label them as

"partitioning" subsets. Otherwise, change the label of *S* into "partitioned."

4. **Stopping partitioning:** If all subsets in the partition scheme ***S*** were labeled "partitioned"; combine all invariant subsets into one subset; and finish the partitioning process and consider ***S*** as the final partition scheme. Otherwise, go to the step 2.

Materials and Experiment Settings

We tested the mPartition algorithm and other partitioning methods on both DNA and protein datasets. The DNA datasets were obtained from the study of RatePartition method including both simulated and real alignments (Rota et al. 2018). The protein datasets included real protein alignments published in previous studies (Ballesteros and Sharma 2019; Chen et al. 2015; Irisarri et al. 2017; Ran et al. 2018; Wu et al. 2018).

Table 1 The real DNA alignments obtained from the RatePartition paper

Alignments	#Taxa	#Loci	#Sites
Arctiina (Katja et al. 2016)	113	8	5809
Calisto (Matos-Maraví 2014)	90	6	5297
Choreutidae (Rota and Wahlberg 2012)	41	8	6293
Coenonymphina (Kodandaramaiah et al. 2009)	69	5	4435
Geometridae (Sihvonen et al. 2011)	164	8	5998
Morpho (Penz, Devries, and Wahlberg 2012)	31	8	6372
Noctuidae (Zahiri et al. 2013)	78	8	6365
Pieridae (Wahlberg et al. 2014)	110	8	6247

Simulated DNA Data

The simulated DNA alignments were generated following 17-leaf trees with both symmetrical and asymmetrical topologies, different branch lengths, nucleotide substitution models, and site rate models (see the RatePartition paper for more details (Rota et al. 2018)). They created 14 sub-datasets each consisted of 20 alignments. Each alignment had four equal partitions of 1000 base pairs simulated under the same substitution model (i.e., F81 or GTR) but with different base frequencies and nucleotide substitution rates. The invariant rate and Gamma distribution rate models were used in the simulations with different proportion of invariant sites and Gamma distribution shapes.

They also created 40 alignments with missing data by randomly removing one of four partitions from some alignments to examine the effect of missing data. In total, the simulated DNA dataset contained 320 alignments.

For each simulated DNA alignment, we evaluated different partition schemes: the true partition scheme with four equal partitions; four partition schemes generated by the RatePartition algorithm with different divide factors

Table 2 The real protein alignments for examining different partitioning methods

Datasets	Clade	Papers	#Taxa	#Loci	#Sites
Ballesteros10	Sea spiders	(Ballesteros and Sharma 2019)	53	10	4046
Ballesteros20	Sea spiders	(Ballesteros and Sharma 2019)	53	20	8575
Ballesteros30	Sea spiders	(Ballesteros and Sharma 2019)	53	30	17045
Ballesteros40	Sea spiders	(Ballesteros and Sharma 2019)	53	40	21998
Chen10	Mammals	(Chen et al. 2015)	58	10	3967
Chen20	Mammals	(Chen et al. 2015)	58	20	6403
Chen30	Mammals	(Chen et al. 2015)	58	30	13267
Chen40	Mammals	(Chen et al. 2015)	58	40	15278
Irissari10	Jawed vertebrates	(Irisarri et al. 2017)	100	10	6027
Irissari20	Jawed vertebrates	(Irisarri et al. 2017)	100	20	8509
Irissari30	Jawed vertebrates	(Irisarri et al. 2017)	100	30	13183
Irissari40	Jawed vertebrates	(Irisarri et al. 2017)	100	40	15698
Ran10	Seed plants	(Ran et al. 2018)	38	10	3062
Ran20	Seed plants	(Ran et al. 2018)	38	20	6897
Ran30	Seed plants	(Ran et al. 2018)	38	30	10443
Ran40	Seed plants	(Ran et al. 2018)	38	40	14749
Wu10	Mammals	(Wu et al. 2018)	90	10	5148
Wu20	Mammals	(Wu et al. 2018)	90	20	12225
Wu30	Mammals	(Wu et al. 2018)	90	30	18088
Wu40	Mammals	(Wu et al. 2018)	90	40	24423

Table 3 The average normalized Robinson–Foulds (RF) distances between the true trees and those constructed with different partition schemes on the simulated DNA alignments

	TruePartition	mPartition	RP2	RP3	RP4	RP5
Average normalized RF distance	0.095	0.115	0.141	0.140	0.141	0.142
Average number of partitions	4	9.6	6.8	9.8	13.2	16.1

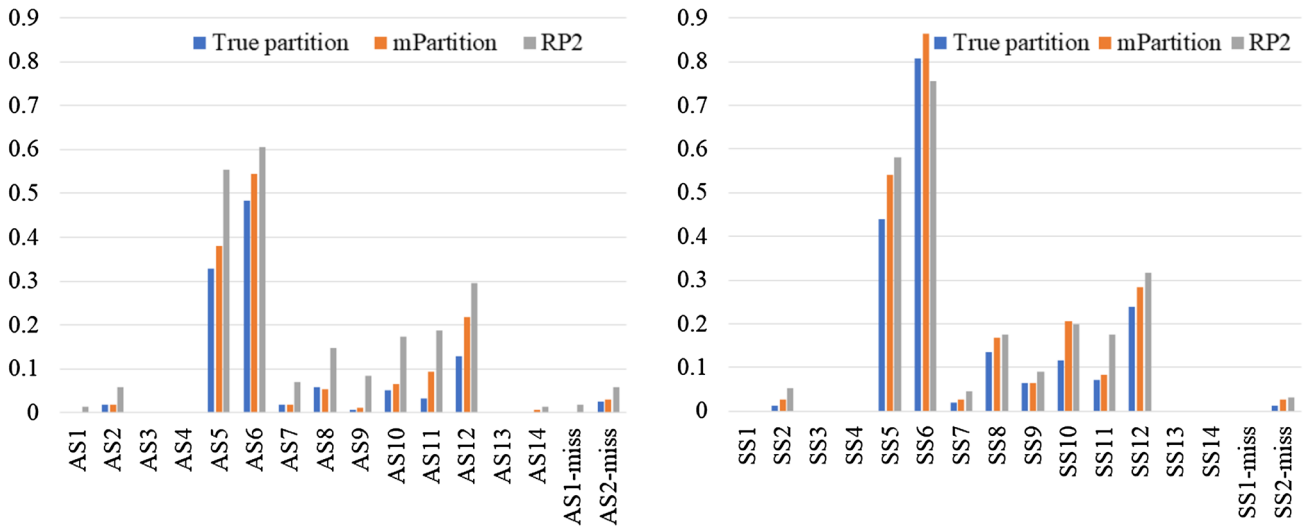

Fig.1 The average normalized Robinson–Foulds distances between the true trees and those inferred from the simulated DNA sub-datasets. *AS* simulations with asymmetrical trees (AS); *SS* simulations with symmetrical trees; miss: missing data; RP2: RatePartition with $d = 2$

$d = 2, 3, 4,$ and 5; and the partition scheme resulted from our mPartition method. For each partition scheme, a ML tree was constructed using IQ-TREE (Minh et al. 2020), and subsequently compared with the topology of the true tree (i.e., the tree used to simulate the alignment) using the normalized Robinson–Foulds (RF) distance (Robinson and Foulds 1981).

Real DNA Data

The real DNA dataset comprised of eight DNA alignments each including one mitochondrial gene (COI) and four to seven nuclear genes commonly used in lepidopteran phylogenetics (see Table 1). The alignment varied in lengths (from 4435 to 6372 sites) and in the number of sequences (from 31 to 164). During the preliminary analyses of these data, the RatePartition algorithm gave better results than other feature-based partition methods based on gene boundaries

Table 4 The AICc and BIC scores of different partitioning methods for the real DNA datasets

Datasets	AICc				BIC			
	NP	RP4	RP5	mPartition	NP	RP4	RP5	mPartition
Arctiina	102,857	101,410	101,225	**99,680**	104,417	103,641	103,533	**101,376**
Calisto	86,492	85,122	85,162	**83,037**	87,733	86,869	87,060	**84,394**
Choreutidae	121,888	115,505	115,381	**113,260**	122,486	117,065	117,211	**114,351**
Coenonymphina	128,991	125,154	125,319	**121,272**	129,946	126,699	126,999	**122,507**
Geometridae	384,948	377,148	377,117	**375,178**	387,217	380,356	380,482	**377,589**
Morpho	58,872	55,862	55,756	**52,633**	59,351	57,095	57,122	**53,539**
Noctuidae	206,920	203,296	203,031	**192,260**	208,080	205,614	205,719	**193,705**
Pieridae	277,253	271,196	271,271	**268,750**	278,805	274,178	274,553	**270,421**

The smaller AICc (BIC) score, the better method. The best AICc and BIC scores are in bold

NP no-partition; *RP4* (*RP5*) RatePartition with $d = 4 (d = 5)$

and/or codon positions on the real DNA alignments (Rota et al. 2018). It was also noted that $d = 4$ and $d = 5$ were the best settings for the RatePartition algorithm for the real DNA dataset. Therefore, we compared 4 partition schemes: mPartition, RatePartition with $d = 4$(RP4) and $d = 5$(RP5), and no-partition. For each partition scheme, a ML tree was constructed using IQ-TREE (Minh et al. 2020). The AICc scores (Hurvich and Tsai 1989) and BIC scores (Schwarz 1978) of constructed trees were used to compare different partition schemes.

Real Protein Data

We compared the partitioning methods on five real concatenated protein alignments obtained from previous studies (see Table 2), i.e., sea spiders, arachnids and several extinct lineage (Ballesteros and Sharma 2019), mammals (Wu et al. 2018), seed pants (Ran et al. 2018), and jawed vertebrates (Chen et al. 2015; Irisarri et al. 2017). The alignments consisted 38 to 100 taxa with thousands of loci. As it was computational burden to examine many different partition schemes on alignments with thousands of loci, we generated smaller alignments concatenated from 10, 20, 30, and 40 randomly selected loci. Note that the 40 loci alignment did not include the 10, 20, or 30 loci alignments. In total, we had 20 real concatenated protein alignments for testing the mPartition and RatePartition methods.

For real protein alignments, the mPartition algorithm selected the best-fit model for a subset from three common general amino-acid substitution models, i.e., LG (Le and Gascuel 2008), JTT (Jones et al. 1992), and WAG (Whelan and Goldman 2001). As the RatePartition algorithm has not been investigated on real protein data, we evaluated its performance with different d values, i.e., $d = 2$(RP2), $d = 3$(RP3), $d = 4$(RP4), and $d = 5$(RP5). We employed IQ-TREE to construct ML trees and used the AICc and BIC scores of constructed trees to compare the goodness of different partition schemes.

Results and Discussions

Simulated DNA Data

We compared normalized RF distances between the true trees and those constructed from true partition schemes (TruePartition), partition schemes created by RatePartition, and partition schemes created by mPartition (see Table 3). Note that the smaller normalized RF distance to the true tree indicates the better partition scheme. Specifically, the average normalized RF distances between the true trees to those constructed with partition schemes from TruePartition, RatePartition, and mPartition were 0.095, 0.141, and 0.115, respectively. Thus, the mPartition method helped infer more accurate trees than the RatePartition method. The small normalized RF distances between trees constructed with the mPartition method and the true trees indicated that mPartition resulted in good partition schemes for the simulated DNA datasets. The RatePartition method performed equally

Table 5 The average normalized Robinson–Foulds distances between partition schemes generated by different partitioning methods

	mPartition	No-partition	RP4	RP5
mPartition	–	0.080	0.076	0.089
No-partition	0.080	–	0.068	0.113
RP4	0.076	0.068	–	0.044
RP5	0.089	0.113	0.044	–

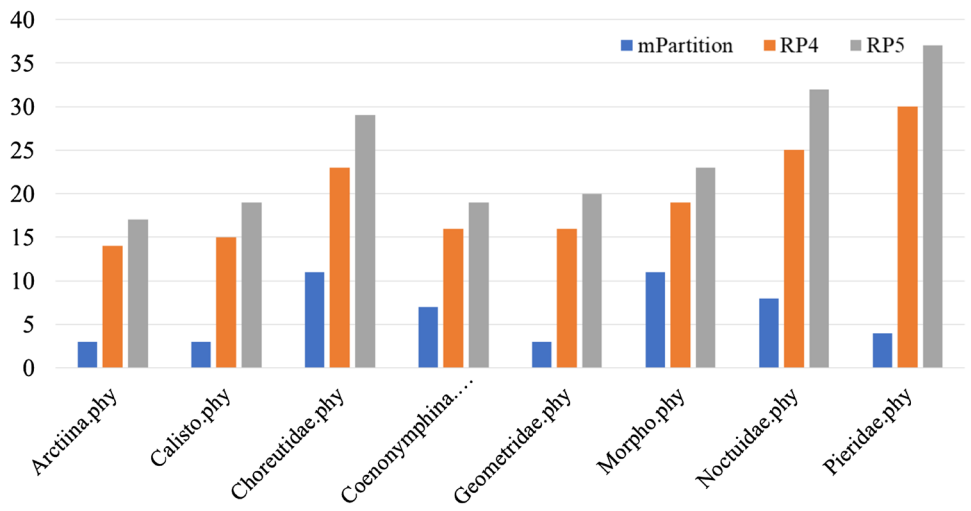

Fig. 2 The sizes of partition schemes generated from different partitioning methods

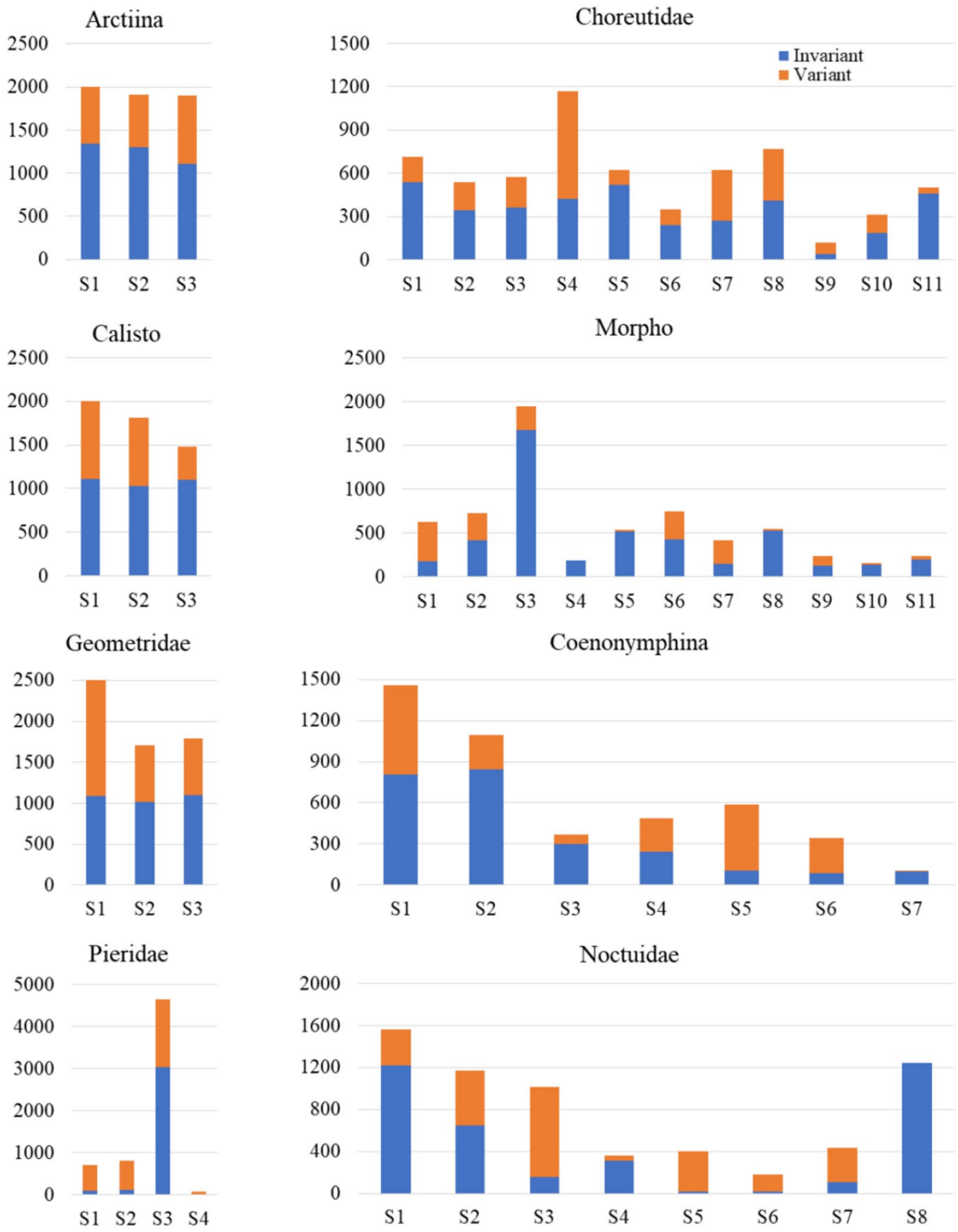

Fig. 3 The distribution of invariant and variant sites in subsets partitioned by the mPartition method

well with different d values (the normalized RF distances were about 0.14 for all RP2, RP3, RP4, and RP5) on the simulated datasets. The mPartition method created about 9.6 subsets per alignment, i.e., more subsets than the RatePartition method with $d = 2$ and less subsets in the other cases.

We also compared the performance of different partitioning schemes on sub-datasets (see Fig. 1). As the RP2, RP3, RP4, and RP5 were equally good on the simulated data, we only analyzed the partition scheme RP2. As anticipated, the performance of partitioning methods was affected by the tree lengths. Specifically, they performed well on sub-datasets (i.e., AS1-AS4, AS13, AS14, SS1-SS4, SS13, SS14), which were simulated based on trees with reasonable long internal branch lengths (≥ 0.01). They had a poor performance on sub-datasets (i.e., AS5, AS6, SS5, SS6), which were generated from trees with very short internal branch

lengths (i.e., 0.001). The tree structures (asymmetrical vs. symmetrical) had some impact on the performance of partitioning methods as the average RF distance of asymmetric inferred trees was lower than that of symmetrical inferred trees. Interestingly, we found that the partitioning methods handled well the missing sub-datasets (AS1-miss, AS2-miss, SS1-miss, SS2-miss).

Real DNA Data

We compared the performance (i.e., in terms of information criteria) of no-partition, RatePartition, mPartition using both AICc and BIC scores (see Table 4). The results indicated that mPartition outperformed other methods in terms of both AICc and BIC scores for all eight real DNA datasets. The no-partition method that did not partition the alignments was the worse method for all alignments. Both AICc and BIC scores of the no-partition method were remarkably worse than that of other partitioning methods. The results suggested that partitioning methods improved the quality of inferred trees. Note that trees constructed with mPartition were different from those constructed with RP4 and RP5 in seven datasets (except the Morpho dataset). The sizes of partition schemes generated by RP4, RP5, and mPartition are presented in Fig. 2. The mPartition method generated less subsets than RP4 and RP5 methods in all datasets.

We computed the average normalized RF distances between inferred trees from different partition schemes (see Table 5). The tree structures constructed with no-partition, mPartition, and RatePartition were different, i.e., the average normalized RF distances between inferred trees with mPartition and those with no-partition, RP4 and RP5 were 0.080, 0.076, and 0.089, respectively. In other words, partition schemes obtained from both mPartition and RatePartition affected tree structures. The distance between trees constructed with RP4 and RP5 (i.e., 0.044) was smaller than those between other partition schemes.

The mPartition method distributed the invariant sites into different subsets, and more importantly a large portion of the subsets were variant sites (see Fig. 3). The invariant sites of the same nucleotide type were partitioned into different subsets. Thus, the mPartition method avoided the pitfall of previous site rate-based partitioning methods, which grouped all invariant sites into one subset.

We also examined the mPartition method on the aquatic beetle family Noteridae dataset, which was used to analyze different partitioning methods (Baca et al. 2017). The mPartition algorithm divided the concatenated alignment into five subsets with sizes from 274 to 2205 sites. The invariants were distributed into all the subsets each containing all four different invariant types (see Fig. 4). We used the UFboot2 algorithm (Hoang et al. 2017) to construct the bootstrap tree from partitioned sub-alignments (see Fig. 5). The mPartition-based bootstrap tree was generally consistent with the bootstrap tree reported in Baca et al. (2017), i.e., all genera were correctly grouped as monophyletic clades with moderate to high support values.

Real Protein Data

We compared the performance (i.e., in terms of information criteria) of the mPartition and RatePartition with $d = 2, 3, 4, 5$ methods on 20 real protein alignments (see Tables 6 and 7). The mPartition method was better than the RatePartition method with different d values in terms of both AICc and BIC criteria. Specifically, the mPartition method was better than the RatePartition methods on 17 out of 20 alignments. Among the partition schemes from RatePartition, RP4 and RP5 performed equally well and better than RP2 and RP3.

We examined the RF distances between trees inferred with different partitioning methods. The average normalized RF distances between trees inferred with mPartition and those inferred with no-partition, RP4, and RP5 were 0.029, 0.032, and 0.034, respectively. The normalized RF distances between trees inferred with different partition schemes from RatePartition were small (i.e., ranging from 0.011 to 0.017).

The mPartition method distributed invariant sites into different subsets, each containing invariant sites of different amino-acid types. For example, Table 8 shows the distribution of invariant sites in the Irissari alignments.

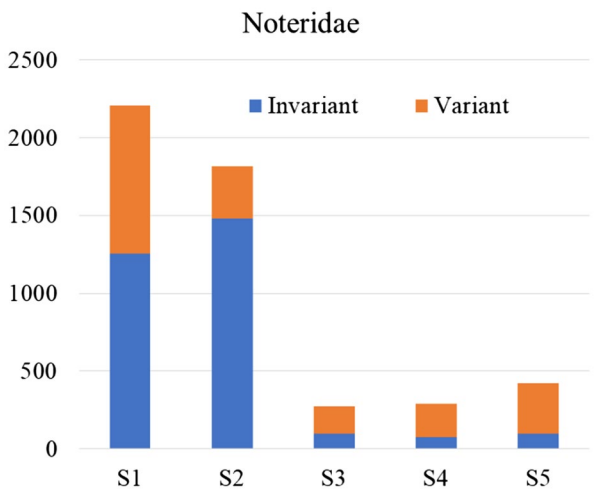

Fig. 4 The distribution of invariant and variant sites in subsets of the Noteridae dataset partitioned by the mPartition method

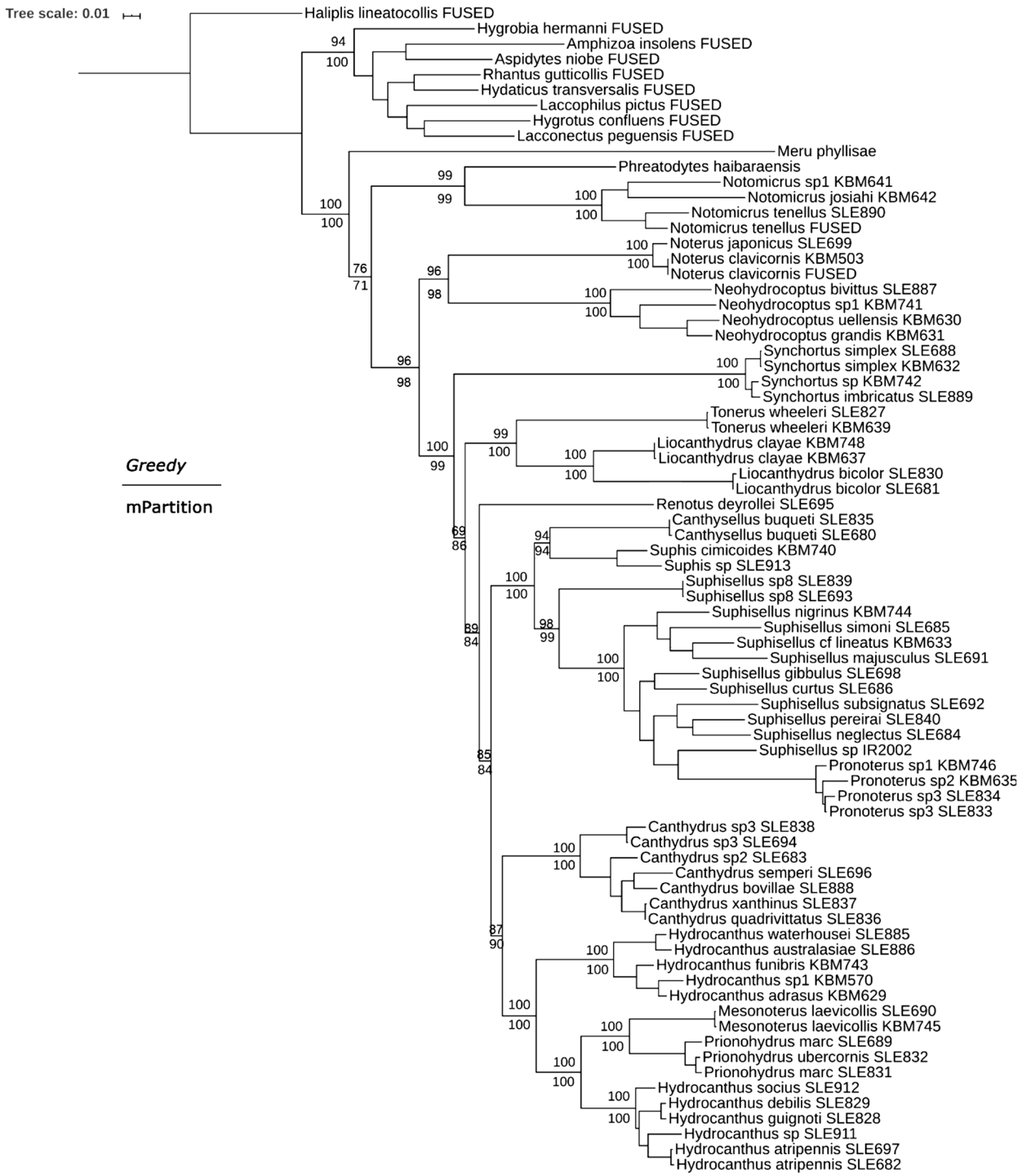

Fig. 5 The bootstrap tree of the aquatic beetle family Noteridae dataset. The tree was constructed by IQ-TREE and the mPartition method. The two numbers X and Y at each branch are the support values of *greedy* method (i.e., partition the alignment based on codon positions and gene fragments) and the mPartition method, respectively

Conclusions

Inferring ML phylogenetic trees from large datasets including multiple loci or whole genomes is a challenging problem due to the complex evolutionary processes among sites. Partitioning alignments into subsets such that sites in the same subset are assumed to evolve under the same evolutionary process has been frequently used in the ML phylogenetic

Table 6 The AICc scores of different partition schemes on the 20 real protein alignments

Datasets	mPartition	RP2	RP3	RP4	RP5
Ballesteros10	**191,471**	195,214	195,194	194,927	195,024
Ballesteros20	**388,710**	392,348	392,096	391,497	391,688
Ballesteros30	**776,551**	789,700	789,219	788,988	788,905
Ballesteros40	**1,156,283**	1,168,122	1,168,039	1,168,049	1,168,168
Chen10	140,998	139,918	139,795	139,581	**139,485**
Chen20	**261,474**	263,092	262,892	264,136	262,680
Chen30	**543,271**	547,649	547,676	547,732	547,589
Chen40	**713,599**	720,539	719,695	719,504	719,491
Irissari10	**319,393**	319,990	320,045	319,949	319,930
Irissari20	**410,737**	416,335	416,108	416,036	416,072
Irissari30	**597,394**	606,780	606,723	606,285	606,482
Irissari40	**765,490**	782,545	782,584	782,565	782,568
Ran10	111,426	**110,952**	111,000	110,962	110,978
Ran20	281,988	281,236	280,911	**280,853**	280,893
Ran30	**409,972**	411,720	411,531	411,625	411,730
Ran40	**593,330**	599,479	596,830	596,778	596,721
Wu10	**189,243**	190,759	190,621	190,572	190,640
Wu20	**586,036**	590,926	590,380	590,211	590,114
Wu30	**755,940**	766,541	766,258	766,046	766,239
Wu40	**1,297,141**	1,304,664	1,304,252	1,303,724	1,303,901

The best AICc scores are in bold. RP2, RP3, RP4, and RP5: RatePartition with $d = 2, 3, 4$ and 5

Table 7 The BIC scores of different partition schemes on 20 real protein alignments

Datasets	mPartition	RP2	RP3	RP4	RP5
Ballesteros10	**192,705**	196,112	196,029	195,972	196,161
Ballesteros20	**389,882**	393,471	393,254	392,746	393,020
Ballesteros30	**779,124**	791,613	791,117	790,771	790,894
Ballesteros40	**1,157,577**	1,169,877	1,170,367	1,170,671	1,170,448
Chen10	141,862	140,732	140,640	140,457	**140,361**
Chen20	**262,655**	263,939	263,786	265,036	263,621
Chen30	**545,438**	548,933	549,049	548,860	548,948
Chen40	**715,922**	721,613	720,875	720,616	720,679
Irissari10	**321,191**	321,566	321,529	321,498	321,513
Irissari20	**412,975**	417,985	417,834	417,776	417,875
Irissari30	**600,039**	608,642	608,504	608,073	608,314
Irissari40	**768,843**	784,612	784,688	784,897	785,029
Ran10	112,285	**111,460**	111,543	111,528	111,585
Ran20	282,892	282,086	281,694	**281,670**	281,750
Ran30	**411,766**	412,645	412,499	412,637	412,641
Ran40	**595,342**	600,450	597,710	597,733	597,698
Wu10	**190,485**	192,026	191,958	191,897	192,028
Wu20	**587,996**	592,526	592,054	591,943	591,985
Wu30	**767,489**	768,338	768,124	768,059	768,206
Wu40	**1,297,295**	1,306,733	1,306,434	1,305,946	1,306,042

The best BIC scores are in bold. RP2, RP3, RP4, and RP5: RatePartition with $d = 2, 3, 4$ and 5

analyses. Partitioning alignments properly is important because it affects both tree topology and branch lengths.

Different computational methods have been proposed to partition alignments based on evolutionary rates of sites. Although the partitioning methods are better than

Table 8 The number of invariant and variant sites in subsets partitioned from the the Irissari alignments by the mPartition method

Irissari 10		Irissari 20		Irissari 30		Irissari 40	
#Invariant	#Variant	#Invariant	#Variant	#Invariant	#Variant	#Invariant	#Variant
696 (20)	1806	36 (10)	385	91 (14)	1030	727 (20)	0
601 (20)	1632	189 (19)	1531	487 (20)	1027	569 (20)	2062
648 (20)	644	282 (20)	301	1650 (20)	1532	287 (20)	1398
		484 (20)	710	476 (20)	2137	533 (20)	497
		405 (20)	328	651 (20)	1122	541 (20)	721
		528 (20)	804	944 (20)	904	407 (20)	1161
		970 (20)	1556	505 (20)	627	667 (20)	561
						127 (17)	652
						27 (6)	823
						1049 (20)	945
						486 (20)	1458

The number of distinct invariant sites in each subset was given in the bracket

no-partition method, the evolutionary rates do not provide sufficient information to comprehensively represent the evolutionary processes of sites. Our mPartition method is designed to partition sites based on the similarity of their site rates as well as substitution models.

Experiments on different datasets showed that mPartition produced better partition schemes than other methods tested. The results on simulated datasets indicated that the trees constructed with mPartition were closer to the true trees than those built with other partitioning methods. The mPartition method also helped building better ML trees than other partitioning methods on both real DNA and protein datasets.

The site rate-based partitioning methods have been widely used, however, they might group invariant sites into one group without any variant sites that might result in incorrect trees. Although the RatePartition method tries to diminish the pitfall by adding some additional variant sites into the subset of invariant sites, our experiments showed that the subset contained only few variant sites. The mPartition algorithm distributed invariant sites into different subsets and more importantly the subsets contained a large number of variant sites to overcome the pitfall of site rate-based methods.

The mPartition algorithm is designed to avoid creating small subsets. Experiments showed that the mPartition algorithm produced less subsets than RatePartition on real datasets. The design is preferred by biologists because they do not have to handle many subsets, site rate models, and substitution models when analyzing real datasets.

Acknowledgements This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.01.2019.06.

References

Baca SM, Toussaint EFA, Miller KB (2017) Molecular phylogeny of the aquatic beetle family Noteridae (Coleoptera: Adephaga) with an emphasis on data partitioning strategies. Mol Phylogenet Evol 107:282–292

Ballesteros J, Sharma P (2019) A critical appraisal of the placement of Xiphosura (Chelicerata) with account of known sources of phylogenetic error. Syst Biol 68:896–917

Brandley MC, Schmitz A, Reeder TW (2005) Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst Biol 54:373–390

Chen MY, Liang D, Zhang P (2015) Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst Biol 64:1104–1120

Crotty SM et al (2020) GHOST: recovering historical signal from heterotachously evolved sequence alignments. Syst Biol 69(2):249–264

Cummins CA, McInerney JO (2011) A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst Biol 60:833–844

Dziak JJ et al (2020) Sensitivity and specificity of information criteria. Brief Bioinform 21(2):553–565

Felsenstein J (2003) Sunderland Inferring Phytogenies. Sinauer Associates, Sunderland

Frandsen PB, Calcott B, Mayer C, Lanfear R (2015) Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol Biol 15:13

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52(5):696–704

Hoang DT et al (2017) UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35(2):518–522

Hurvich CM, Tsai C-L (1989) Regression and time series model selection in small samples. Biometrika 76:297–307

Irisarri I et al (2017) Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat Ecol Evol 1:1370–1378

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Bioinformatics 8(3):275–282

Kainer D, Lanfear R (2015) The effects of partitioning on phylogenetic inference. Mol Biol Evol 32:1611–1627

Kalyaanamoorthy S et al (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587–589

Katja R, Mappes J, Lauri K, Wahlberg N (2016) Putting Parasemia in its phylogenetic place: a molecular analysis of the subtribe Arctiina (Lepidoptera): molecular phylogeny of Arctiina. Syst Entomol 41:844–853

Kodandaramaiah U et al (2009) Phylogenetics of Coenonymphina (Nymphalidae: Satyrinae) and the problem of rooting rapid radiations. Mol Phylogenet Evol 54:386–394

Kumar S et al (2012) Statistics and truth in phylogenomics. Mol Biol Evol 29:457–472

Lanfear R, Calcott B, Ho S, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695–1701

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109

Le SQ, Dang CC, Gascuel O (2012) Modeling protein evolution with several amino-acid replacement matrices depending on site rates. Mol Biol Evol 29:2921–2936

Le SQ, Gascuel O (2008) An improved general amino-acid replacement matrix. Mol Biol Evol 25(7):1307–1320

Lemey P, Salemi M, Vandamme AM (2009) The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing, 2nd edn. Cambridge University Press, Cambridge

Matos-Maraví P et al (2014) Causes of endemic radiation in the Caribbean: Evidence from the historical biogeography and diversification of the butterfly genus Calisto (Nymphalidae: Satyrinae: Satyrini). BMC Evol Biol 14:199

Minh BQ et al (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37(5):1530–1534

Nylander J, Ronquist F, Huelsenbeck J, Nieves-Aldrey J (2004) Bayesian phylogenetic analysis of combined data. Syst Biol 53:47–67

Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571–581

Penz C, Devries P, Wahlberg N (2012) Diversification of Morpho butterflies (Lepidoptera, Nymphalidae): a re-evaluation of morphological characters and new insight from DNA sequence data. Syst Entomol 37:670–685

Ran J-H, Shen T-T, Wang M-M, Wang X-Q (2018) Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between gnetales and angiosperms. Proc Royal Soc B. https://doi.org/10.1098/rspb.2018.1012

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147

Rodréguez-Ezpeleta N et al (2007) Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol 56:389–399

Rota J et al (2018) A simple method for data partitioning based on relative evolutionary rates. PeerJ 6:e5498

Rota J, Wahlberg N (2012) Exploration of data partitioning in an eight-gene data set: phylogeny of metalmark moths (Lepidoptera, Choreutidae). Zoologica Scripta 41(5):536–546

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Sihvonen MP et al (2011) Comprehensive molecular sampling yields a Robust Phylogeny for Geometrid Moths (Lepidoptera: Geometridae). PLoS One 6:e20356

Stamatakis A (2015) Using RAxML to Infer Phylogenies. Curr Protoc Bioinformatics 51:6.14.1–6.14.14

Strimmer K, von Haeseler A (1997) Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci USA 94(13):6815–6819

Tagliacollo V, Lanfear R (2018) Estimating improved partitioning schemes for ultraconserved elements (UCEs). Mol Biol Evol 35(7):1798–1811

Vinh LS, von Haeseler A (2004) IQPNNI: Moving fast through tree space and stopping in time. Mol Biol Evol 21(8):1565–1571

Wahlberg N et al (2014) Revised systematics and higher classification of pierid butterflies (Lepidoptera: Pieridae) based on molecular data. Zool Scr 43:641–650

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18:691–699

Wu S, Edwards S, Liang L (2018) Genome-scale DNA sequence data and the evolutionary history of placental mammals. Data Brief 18:1972–1975

Yang Z. (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10(6):1396–1401

Yang Z (1996) Maximum-likelihood models for combined analyses of multiple sequence data. J Mol Evol 42(5):587–596

Zahiri Z et al (2013) Relationships among the basal lineages of Noctuidae (Lepidoptera, Noctuoidea) based on eight gene regions. Zool Scr 42:488–507