HOKKAIDO UNIVERSITY

Title	PsiPartition : Improved Site Partitioning for Genomic Data by Parameterized Sorting Indices and Bayesian Optimization
Author(s)	Xu, Shijie; Onoda, Akira
Citation	Journal of Molecular Evolution, 92(6), 874-890 https://doi.org/10.1007/s00239-024-10215-7
Issue Date	2024-12-05
Doc URL	https://hdl.handle.net/2115/99118
Rights	This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s00239-024-10215-7
Type	journal article
File Information	PsiPartition_ShijieXu_Manuscript_20241117-3.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

PsiPartition: Improved Site Partitioning for Genomic Data by Parameterized Sorting Indices and Bayesian Optimization

Shijie Xu[1] and Akira Onoda[1,2*]

[1]Graduate School of Environmental Science, Hokkaido University, Kita-10 Nishi-5, Kita-ku, Sapporo, 060-0810, Hokkaido, Japan.
[1]Faculty of Environmental Earth Science, Hokkaido University, Kita-10 Nishi-5, Kita-ku, Sapporo, 060-0810, Hokkaido, Japan.

*Corresponding author(s). E-mail(s): akira.onoda@ees.hokudai.ac.jp;
Contributing authors: shijie.xu@ees.hokudai.ac.jp;

Abstract

Phylogenetics has been widely used in molecular biology to infer the evolutionary relationships among species. With the rapid development of sequencing technology, genomic data with thousands of sites becomes increasingly common in phylogenetic analysis, while heterogeneity among sites arises as one of the major challenges. A single homogeneous model is not sufficient to describe the evolution of all sites and partitioned models are often employed to model the evolution of heterogeneous sites by partitioning them into distinct groups and utilizing distinct evolutionary models for each group. It is crucial to determine the best partitioning, which greatly affects the reconstruction correctness of phylogeny. However, the best partitioning is usually intractable to obtain in practice. Traditional partitioning methods rely on heuristic algorithms or greedy search to determine the best ones in their solution space, are usually time-consuming, and with no guarantee of optimality. In this study, we propose a novel partitioning approach, termed PsiPartition, based on the parameterized sorting indices of sites and Bayesian optimization. We apply our method to empirical data sets and it performs significantly better compared to existing methods, in terms of Bayesian information criterion (BIC) and the corrected Akaike information criterion (AICc). We test PsiPartition on the simulated data sets with different site heterogeneity, alignment lengths, and number of loci. It is demonstrated that PsiPartition evidently and stably outperforms other methods in terms of the Robinson-Foulds (RF) distance between the true simulated trees and the reconstructed trees, especially on the data with more site heterogeneity. More

importantly, our proposed Bayesian optimization-based method, for the first time, provides a new general framework to efficiently determine the optimal number of partitions. The corresponding reproducible source code and data are available at http://github.com/xu-shi-jie/PsiPartition.

Keywords: Phylogenetic, Phylogenomics, Bayesian optimization, Partitioned models

1 Introduction

Phylogenetics provides powerful tools to reconstruct the evolutionary history of organisms based on their genetic sequences. In the past decades, various methods have been proposed to reconstruct phylogenetic trees such as the neighbor joining Saitou and Nei (1987), maximum-likelihood (ML) Felsenstein (1981), Bayesian inference (BI) Rannala and Yang (1996), and graph splitting Matsui and Iwasaki (2020). ML and BI are generally considered as the most accurate phylogenetic methods by modeling the evolution process as a continuous-time Markov process with a specific substitution model Yang (2006). With the rapid development of sequencing technology, the size of the data sets has been increasing dramatically. Phylogenetics on genomic data, named phylogenomics, has become a hot topic in recent years. For example, DNA sequences from distinct genes of the species are often selected and concatenated into a supermatrix, to infer the phylogeny of species more accurately Chernomor et al (2016). It has been suggested phylogeny inferred based on more sites is better than fewer Rosenberg and Kumar (2003). However, the heterogeneity among sites becomes one of the major challenges in phylogenomic analysis. Genes in different regions of the genome may evolve at different rates, e.g., the third codon position of protein-coding genes evolves faster than the first and second codon positions Bofkin and Goldman (2007) because the third codon position is often under less selective pressure. In addition, the heterogeneity of amino acids may arise from the different secondary structures of proteins Le and Gascuel (2010). Such heterogeneity is ubiquitous in genomic data and a single homogeneous model is not sufficient to describe it. This limits the potential improvements in the accuracy of phylogenetic inference by enhancing evolutionary signals with multiple genes Kainer and Lanfear (2015). Alternative methods that reduce the influence of site heterogeneity among sites are thus needed.

Partitioned models are proposed to describe the site heterogeneity and improve the accuracy of phylogenetic inference. In partitioned models, sites are divided into several disjoint subsets that apply different models, each of which may have different substitution rates, base frequencies, and branch lengths Lanfear et al (2017). The importance of partitioned models has been demonstrated in phylogenomic analysis Chernomor et al (2016), as they can affect tree branches, topology, bootstrap support, and divergence date Brandley et al (2005); Ho and Lanfear (2010); Leavitt et al (2013); Poux et al (2008); Rota and Wahlberg (2012); Le Kim and Le Sy (2020). Many parameters of partitioned models affect the accuracy of phylogenetic inference and one of the most essential factors is the number of partitions. In general, more partitions can fit the data better as they utilize more evolutionary models. However, they also increase

the risk of overfitting, which biases the subsequent analysis Frandsen et al (2015). It is crucial to balance the goodness of fit and the complexity of the models. Information criterion such as the Akaike information criterion (AIC) Akaike (1974), corrected AIC (AICc) Sugiura (1978), and Bayesian information criterion (BIC) Schwarz (1978) are often employed as an optimized objective in partitioned models.

There remain critical problems in partitioning sites efficiently and properly. Firstly, the number of all possible partitioning of an n-sites alignment is the Bell number B_n satisfying the recursive definition $B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k$, which increases faster than an exponential function as n increases. It is impossible to perform a brute-force search to exhaust all partitioning schemes even for alignments with a small number of sites Lanfear et al (2012). Secondly, partitioning methods using *prior* knowledge such as codon positions and secondary structures of proteins are based on intuition, do not guarantee a correct partitioning of sites, and have therefore been improved in a series of works Brandley et al (2005); Brown and Lemmon (2007); McGuire et al (2007); Li et al (2008). Their methods partition sites based on a handful selection of collections of partitioning candidates, which requires expert knowledge and may lead to bias. Automatic partitioning methods have been proposed to search for a good partition by heuristic algorithms in the whole partitioning space. For example, the reported work Frandsen et al (2015) proposed a top-down greedy algorithm that partitions the sites from a single partition into more partitions gradually. It is based on the k-means clustering MacQueen et al (1967) of the site rates estimated from the reconstructed phylogenetic trees. Instead of partitioning sites, the reported work Lanfear et al (2014) recursively merges the most similar site pairs. These methods have achieved excellent performance while still having some disadvantages: 1) To acquire the estimated site rates, they rely on reconstructed reference phylogenetic trees, which take time to obtain and are possibly incorrect, and 2) they are based on greedy search and could fall in the local optimum. These disadvantages limit their applications in the situation with genomic data.

Recently, TIGER (Tree Independent Generation of Evolutionary Rates) has been proposed to estimate the evolutionary rates of sites without referring to phylogenetic trees Cummins and McInerney (2011). It was originally developed to exclude the fast-evolving sites that interfere with the accurate reconstruction of actual phylogeny. TIGER-based partitioned models have demonstrated their success in many phylogenetic studies Kaila et al (2013); Rota and Miller (2013); Heikkilä et al (2014); Matos-Maraví et al (2014); Wahlberg et al (2014); Edger et al (2015); Kristensen et al (2015); Rajaei et al (2015); Ounap et al (2016). Rota *et al.* attempted to improve TIGER by using an extra "division factor" defined by users Rota et al (2018). Kim *et al.* made efforts to automatically select the best number of partitions by recursive trisection Le Kim and Le Sy (2020). These two methods alleviate the demands of automatic partitioning for genomic data. However, current TIGER-based methods still do not achieve the best accuracy. More accurate and efficient partitioning algorithms are thus needed.

In this study, we propose a novel site partitioning approach, termed PsiPartition, based on Parameterized Sorting Indices (PSI) and Bayesian optimization. We first

introduce the PSI as a better alternative to TIGER in terms of accuracy with reasonable computational costs. Then we optimize PSI and the number of partitions by Bayesian optimization. We demonstrate that our method significantly outperforms the existing methods on empirical data sets in terms of Bayesian information criterion and corrected Akaike information criterion. Furthermore, it reconstructs the most accurate phylogenetic trees on the simulated data sets, in terms of Robinson-Foulds distance. More importantly, our method provides a new general Bayesian optimization-based framework to efficiently determine the best number of partitions and can be easily extended to other partitioning methods.

2 Method

2.1 Data sets

For phylogenetic inference benchmarking, we employ eight empirical DNA data sets from the reported work Rönkä et al (2016); Matos-Maraví et al (2014); Rota and Wahlberg (2012); Kodandaramaiah et al (2010); Sihvonen et al (2011); Penz et al (2012); Zahiri et al (2013); Wahlberg et al (2014). Their statistics are listed in Table 1. These data sets are also employed as the benchmark data sets in previous studies Le Kim and Le Sy (2020); Rota et al (2018). Alignments in the data sets have a range of base pair lengths from 4435 to 6372 and the number of taxa from 31 to 164. Each sequence in the alignments is the concatenated sequences from the mitochondrial gene (COI) and other four to seven nuclear genes commonly employed for lepidopteran phylogenetics, including CAD, EF-1 α, GAPDH, IDH, MDH, RpS5, and wingless Wahlberg and Wheat (2008).

Table 1 The empirical DNA data sets for benchmarking.

Name	No. of taxa	No. of loci	No. of sites
Arctiina Rönkä et al (2016)	113	8	5809
Calisto Matos-Maraví et al (2014)	90	6	5297
Choreutidae Rota and Wahlberg (2012)	41	8	6293
Coenonymphina Kodandaramaiah et al (2010)	69	5	4435
Geometridae Sihvonen et al (2011)	164	8	5998
Morpho Penz et al (2012)	31	8	6372
Noctuidae Zahiri et al (2013)	78	8	6365
Pieridae Wahlberg et al (2014)	110	8	6247

Six empirical protein data sets from the reported work Nguyen et al (2016) are also employed for benchmarking on the protein data. Their statistics are listed in Table 2. Alignments in the data sets have a range of amino acid lengths from 385 to 811 and the number of taxa from 17 to 31. Each alignment in the data sets includes the sequences of heat shock proteins Hsp83, Hsc70-3, Hsc70-4, Hsc70-5, Hsp60, and Hsp40. These proteins are the primary molecular chaperones protecting against thermal damage. Note that they are relatively small proteins and the number of sites is smaller compared to the DNA data sets.

Table 2 The empirical protein data sets for benchmarking.

Name	No. of taxa	No. of sites
Nguyen_2016a	17	688
Nguyen_2016b	31	680
Nguyen_2016c	25	811
Nguyen_2016d	17	704
Nguyen_2016e	17	385
Nguyen_2016f	17	583

2.2 Parameterized Sorting Indices (PSI)

We introduce the proposed Parameterized Sorting Indices (PSI) by reviewing the Tree-Independent Generation of Evolutionary Rates (TIGER) Cummins and McInerney (2011). In the original paper of TIGER, the evolutionary rates of site i are estimated as

$$r_i = \frac{1}{n-1} \sum_{j \neq i} \text{pa}(i,j) \tag{1}$$

where n is the number of sites and

$$\text{pa}(i,j) = \frac{1}{|P_j|} \sum_{x \in P_j} a(x, P_i)$$

is the *agreement score*. Here the function $a(x, P_i)$ is defined as $a(x, P_i) = 1$ if $x \subset A$ and $A \in P_i$, and otherwise 0. The set P_i is the *set partition* of the site i, whose elements are disjoint subsets of $\{1, 2, ..., m\}$ where m is the number of taxa. For each subset $A \in P_i$, the character of the l-th sequence at site i has the same type for all $l \in A$. For example, consider two sites S_i = CTTAA and S_j = AGGGG with the set partition $P_i = \{\{1\}, \{2,3\}, \{4,5\}\}$ and $P_j = \{\{1\}, \{2,3,4,5\}\}$, respectively. We have $\text{pa}(A,B) = 0.5$ because out of two subsets $\{1\}$ and $\{2,3,4,5\} \in P_j$, only $\{1\} \subset \{1\} \in P_i$. $\text{pa}(B,A) = 1$ because for all subsets $\{1\}, \{2,3\}, \{4,5\} \in P_i$, we have $\{1\} \subset \{1\} \in P_j$, $\{2,3\} \subset \{2,3,4,5\} \in P_j$, and $\{4,5\} \subset \{2,3,4,5\} \in P_j$. The TIGER of site S_i and S_j should be $r_i = 3/3 = 1$ and $r_j = 1/3$, respectively.

PSI has two key innovations. Firstly, in the original TIGER, the function $a(x, P_i)$ is defined as binary values 1 or 0 for all characters x at site j, regardless of the characters' types. However, characters representing different base pairs or amino acids should have different properties and should be treated differently. This is similar to the idea of base frequency parameters implemented in most phylogenetic inference tools Nguyen et al (2015); Höhna et al (2016). We therefore introduce a group of parameters w to represent these differences among characters. For example, we adopt five parameters $w = \{w_A, w_C, w_G, w_T, w_{\text{gap}}\}$ for DNA data sets, where w_A is the parameter for character A, w_{gap} is the parameter for character gaps or missing characters, and so on. We define $a_w(x, P_i) = w_{p(x)}$ if $x \subset A$ and $A \in P_i$ and otherwise 0 where $p(x)$ is the corresponding character of subset x. Similarly, there should be 21 parameters for protein data sets.

Secondly, instead of using Equation 1, we utilize a comparison between two sites to define PSI at site i by following

$$\mathrm{PSI}_i = |\{j|\mathrm{pa}''(i,j) > \mathrm{pa}''(j,i)\}|$$

where $\mathrm{pa}''(i,j)$ is the defined as following

$$\mathrm{pa}''(i,j) = \mathrm{pa}'(i,j) + \delta_{ij}, \quad \mathrm{pa}'(i,j) = \frac{1}{|P_j|}\sum_{x\in P_j} a_w(x, P_i).$$

It is crucial to avoid grouping all invariant sites into the same partition, since this may increase the likelihood but result in a biased phylogeny, as highlighted by previous studies Baca et al (2017); Rota et al (2018); Le Kim and Le Sy (2020). In site rate-based partitioning methods, invariant sites with identical characters in an alignment are indistinguishable from one another and are often grouped into the same partition. We address this indistinguishability by introducing additional noise, modeled as a uniform distribution, $\delta_{ij} \sim \mathrm{U}(0, m)$, where $m = \frac{1}{n^2}\sum_{i,j} \mathrm{pa}'$ is the mean of pa' over all pairs of sites. Amino acid sequence are more likely to be variable than DNA sequences, since the former has 20 characters while the latter has only 4. We therefore do not employ the noise term by setting $\delta_{ij} = 0$ for protein data sets. The PSI is a non-negative integer ranging from 0 to $n-1$ represents the number of sites that site i agrees more than it is agreed by.

After calculating the PSI for all sites, we bin the sites into k partitions based on their PSIs. Denote $\alpha = \min_i \mathrm{PSI}_i$ and $\beta = \max_i \mathrm{PSI}_i$, then k partitions are defined as following

$$\begin{aligned}
P_1 &= \{i|\mathrm{PSI}_i \in [\alpha, \alpha + \frac{\beta-\alpha}{k})\} \\
P_2 &= \{i|\mathrm{PSI}_i \in [\alpha + \frac{\beta-\alpha}{k}, \alpha + 2\frac{\beta-\alpha}{k})\} \\
&\dots \\
P_k &= \{i|\mathrm{PSI}_i \in [\alpha + (k-1)\frac{\beta-\alpha}{k}, \beta]\}
\end{aligned}$$

where k is the number of partitions. For example, if $\alpha = 0$ and $\beta = 10$, then 3 partitions are $P_1 = \{i|\mathrm{PSI}_i \in [0, \frac{10}{3})\}$, $P_2 = \{i|\mathrm{PSI}_i \in [\frac{10}{3}, \frac{20}{3})\}$, and $P_3 = \{i|\mathrm{PSI}_i \in [\frac{20}{3}, 10]\}$.

Note that the parameters w and the number of partitions k described in this section are not determined and will be optimized based on the input data.

2.3 Bayesian optimization for PSI

It is important to decide the best number k of partitions in the partitioned models. When k equals the number of sites, a single-site partitioning can fit the data very well. This may cause overfitting and produce the reconstructed phylogeny different from the true one. When $k = 1$, the single partition is not suitable for describing heterogeneity among the sites. Some existing methods employ greedy algorithms to determine the number of partitions which is time-consuming for large data sets.

Bayesian optimization (BO), a black-box method, has been widely used in many fields for parameter tuning and are also adopted in this study. It has a significant advantage when the objective function is expensive to evaluate and the number of parameters is small. In our situation, the objective function is the information criterion (i.e., BIC) of the partitioned models, which is expensive to evaluate because it requires full phylogenetic inference on the given data. The optimized parameters are the number of partitions k and the introduced parameters w of PSI, which are few (i.e., 6 for DNA and 22 for proteins). BO utilizes a surrogate model such as the Gaussian process model Williams and Rasmussen (1995) to approximate the objective function. The parameters are estimated based on the acquisition functions such as expectation improvement and upper confidence bound. Then BO evaluates the objective function on these parameters to obtain a new observation and updates the surrogate model with all observations. This process repeats until the parameter converges or the maximum number of iterations is reached.

We propose PsiPartition algorithms based on BO, as shown in Algorithm 1. The input of the algorithm is the alignment A, the maximum number of partitions $k_{\max}$, and the maximum iterations steps n. The output is the optimal partition P^*, the optimal parameters s^* and the optimal objective value o^*. The algorithm firstly initializes an empty set S and samples the number of partitions k_1 from $\{2, 3, ..., k_{\max}\}$ and the parameters w_1 from $[0, 1]^t$ where $t = 5$ for DNA data and $t = 21$ for amino acid data. For each iteration i, it calculates the PSI ψ_i of all sites with parameters w_i and bins the sites into k_i partitions $P_i = \{p_j\}_{j=1}^{k_i}$ based on ψ_i. It thus evaluates the phylogenetic software (e.g., IQ-TREE) with the partitioning P_i to obtain the BIC b_i. Each evaluation is recorded as an observation (P_i, s_i, b_i) where $s_i = (w_i, k_i)$, and pushed into the set S. Then it updates the surrogate model m with all observations $(P_i, s_i, b_i) \in S$ and determines the next parameters $s_{i+1} = (w_{i+1}, k_{i+1})$ by the surrogate model m with expected improvement. The iterations will be repeated for n times. Finally, the algorithm returns the partitioning P^*, the parameters s^*, and the BIC o^* that has the minimum BIC o^* in S.

To indicate the necessity of BO, we develop a variant of PsiPartition, termed PsiPartitionFast, which freezes all parameters w_i as 1 in Algorithm 1. Compared to PsiPartition, it only optimizes the number of partitions k. The search space can be reduced and the overall steps of optimization are also less than PsiPartition. A detailed description is referred to the pseudocode in Supplementary Algorithm S1.

2.4 Evaluation metrics

To evaluate the performance of different partitioning methods, we employ the Bayesian information criterion (BIC) Schwarz (1978) and the corrected Akaike information criterion (AICc) Sugiura (1978). The BIC and AICc are defined as the following

$$\mathrm{BIC} = -2 \ln L + k \ln n$$

and

$$\mathrm{AICc} = -2 \ln L + 2k + \frac{2k(k+1)}{n-k-1}$$

Algorithm 1 PsiPartition

Require:

A: the alignment

$k_{\max}$: the maximum number of partitions

n: the maximum iteration steps

Ensure:

P^*: the optimal partition

s^*: the optimal parameters

o^*: the optimal objective value

Uniformly sample k_1 from $2, 3, ..., k_{\max}$

Uniformly sample w_1 from $[0, 1]^t$ ▷ $t = 5$ for DNA, 21 for protein

Let S be an empty set

Initialize a surrogate model m

for $i = 1$ to n **do**

    Let ψ_i be the PSI of all sites with parameters w_i

    Bin the sites of A into k_i partitions based on ψ_i as P_i

    Evaluate phylogenetic tool (e.g., IQ-TREE) with partitioning P_i and obtain BIC b_i

    $s_i \leftarrow (k_i, w_i)$

    $S \leftarrow S \cup \{(P_i, s_i, b_i)\}$

    Update m with S

    Determine the next parameters $s_{i+1} = (k_{i+1}, w_{i+1})$ by the surrogate model m with expected improvement

end for

return $(P^*, s^*, o^*) \in S$ that has the minimum BIC o^*

where L is the likelihood of the model, k is the number of model parameters, and n is the number of sites. BIC and AICc are both used to assess the goodness of partitioned models, which with the smallest BIC or AICc are considered the best ones. The BIC is the default criterion for model selection in IQ-TREE software Kalyaanamoorthy et al (2017). We thus adopted it as the major evaluation metric and AICc as a reference. It is worth noting that the partitioning of sites is usually considered as *a priori*. Therefore the parameters w introduced in the aforementioned Bayesian optimization algorithm and the number of partition k should not be included in the calculation of BIC and AICc.

The Robinson-Foulds (RF) distance Robinson and Foulds (1981) is adopted to evaluate the accuracy of the reconstructed phylogenetic trees. Two branches sharing a common ancestor in a tree split a part of taxa into two disjoint subsets (i.e., clades). Therefore they recursively imply a nested partitioning of the taxa. The RF distance is defined as the summation of the number of partitions implied by one tree but not by another one. It is a metric of the topological distance between two trees, and the smaller the RF distance, the more similar the two trees. In our experiments, we compute the RF distance between the true simulated trees and the reconstructed trees and compare them of different partitioning methods.

We utilize the transfer distance to compare the similarity between two partitions. Suppose X be a set, and $P = \{C_1, C_2, ..., C_n\}$ and $P' = \{C'_1, C'_2, ..., C'_m\}$ are two partitions of X. Without loss of generality, we assume $n = m$ because we can add empty subsets to the smaller one. We firstly define the *concordance* as following

$$c(P, P') = \max_{\sigma \in M} \sum_{i=1}^{n} |C_i \cap C'_{\sigma(i)}|$$

where M is the set of all permutations of $\{1, 2, ..., n\}$. The *transfer distance* is defined as following

$$t(P, P') = 1 - \frac{c(P, P')}{|X|}.$$

It is obvious to see that $0 \leq t(P, P') \leq 1$ and $t(P, P') = 0$ if and only if $P = P'$. A smaller transfer distance means more similarity between two partitions. The proof of the properties of the transfer distance is described in Day (1981) with more details.

3 Results

PsiPartition is implemented within Python 3.12.1. To perform the Bayesian optimization, wandb package (version 0.16.1) Biewald et al (2020) are also used Wandb (2024). It employed normalized Gaussian process model with Matern kernel Rasmussen et al (2006) and parameter $\alpha = 10^{-7}$. Expected improvement is utilized as the default acquisition function. The parameters $k_{\max}$ and n in Algorithm 1 are both set up to 30 by default unless mentioned. The partitioned models and their information criterion are evaluated by ModelFinder Kalyaanamoorthy et al (2017) implemented in IQ-TREE Nguyen et al (2015) (version 1.6.12). UFBoot Minh et al (2013) in IQ-TREE is also enabled for ultrafast bootstrap calculation of phylogenetic trees. The experiments are performed on a Linux machine with an AMD Ryzen 9 7950X3D 16-core processor and 64 GB memory. Multiple threads feature of IQ-TREE are adopted to accelerate the phylogenetic inference. Interactive Tree Of Life (ITOL) online tools Letunic and Bork (2021) are used to visualize the phylogenetic trees and the bootstrap support values of branches.

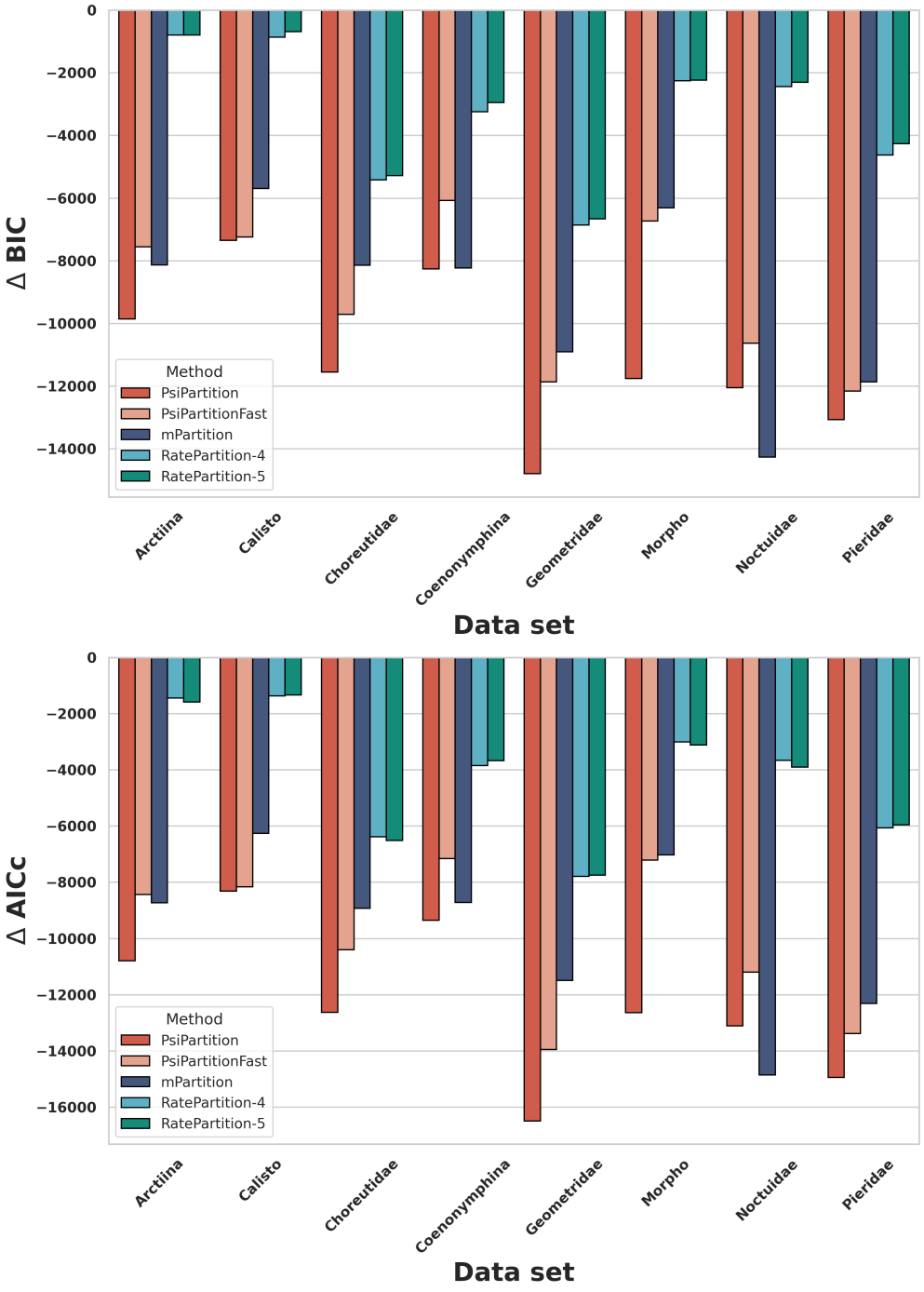


Fig. 1 The performance comparison of different partitioning methods on the empirical DNA data sets (Arctiina, Calisto, Choreutidae, Coenonymphina, Geometridae, Morpho, Noctuidae, and Pieridae). Our proposed methods (PsiPartition and PsiPartitionFast) achieves the best performance in terms of the ΔBIC and ΔAIC, significantly outperforming the other methods.

3.1 PsiPartition outperforms other partitioning methods on empirical data sets

To evaluate the performance of our proposed methods, PsiPartition and PsiPartitionFast, and existing state-of-the-art methods, we applied them to eight empirical DNA data sets listed in Table 1. The experimental results are shown in Fig 1. The compared methods include the default IQ-TREE using the single homogeneous model without partitioning (NP), RatePartition with parameter div_factor = 4 (RatePartition-4) and div_factor = 5 (RatePartition-5), and mPartition. We show the advantages of partitioned models compared to NP by BIC decreases defined as the following

$$\begin{aligned}\Delta\text{BIC} = {} & \text{BIC of partitioned models} - \\ & \text{BIC of model without partitioning},\end{aligned}$$

and ΔAICc is defined similarly.

On all data sets, our proposed methods PsiPartition and PsiPartitionFast significantly outperform other partitioned methods in terms of the ΔBIC and ΔAICc. Both of two significantly reduce the BIC and AICc values of the partitioned models, as shown in Supplementary Table S1. In addition, PsiPartition is better than PsiPartitionFast, which suggests that the parameters w play a vital role for partitioning. On the data set Calisto, PsiPartitionFast is slightly better than PsiPartition, which is probably caused by the relatively small size of the data set and the randomness in phylogenetic inference Kumar et al (2023). The results demonstrate that our proposed methods are more accurate than the default IQ-TREE and other existing partitioning methods, especially on the data sets with more sites and taxa.

We visualize the distribution of partitions in Fig 2. Considering the differences among data, the number of partitions k determined by PsiPartition for different empirical DNA data sets are distinct, ranging from 17 to 29. This suggests that our proposed methods are capable of searching for different optimal partition schemes on different data sets. Furthermore, the number of variant and invariant sites that fall into each partition are also plotted. Bin indices represent the binning of sites based on the PSI, i.e., the sites with lower PSI will be assigned to the smaller bin indices, and vice versa. It is clear that the distribution of the number of sites in each bin is quite imbalanced and most of the sites are assigned to the first several bins, while there are still a few sites assigned to the last several bins. It is also worth noting that invariant sites are distributed in the several different bins and the first several bins contain slow-evolving variant sites as well, suggesting our method avoids grouping all invariant sites into the same partition. This is similar to those in previous study and reveals that our methods properly partition the sites.

We further illustrate that the noise term is necessary for the PSI on the DNA data sets. As a comparison, we apply the PsiPartition without the noise term to estimate the site rates and the distribution of invariant and variant sites are shown in Supplementary Fig. S2. Though PSI without noise terms generate with grouping of invariant sites into the same partition for most data sets, Pieridae distribute almost all invariant sites into the first partition. To increase the robustness of the PSI, we suggest that the noise term is necessary for the DNA data sets.

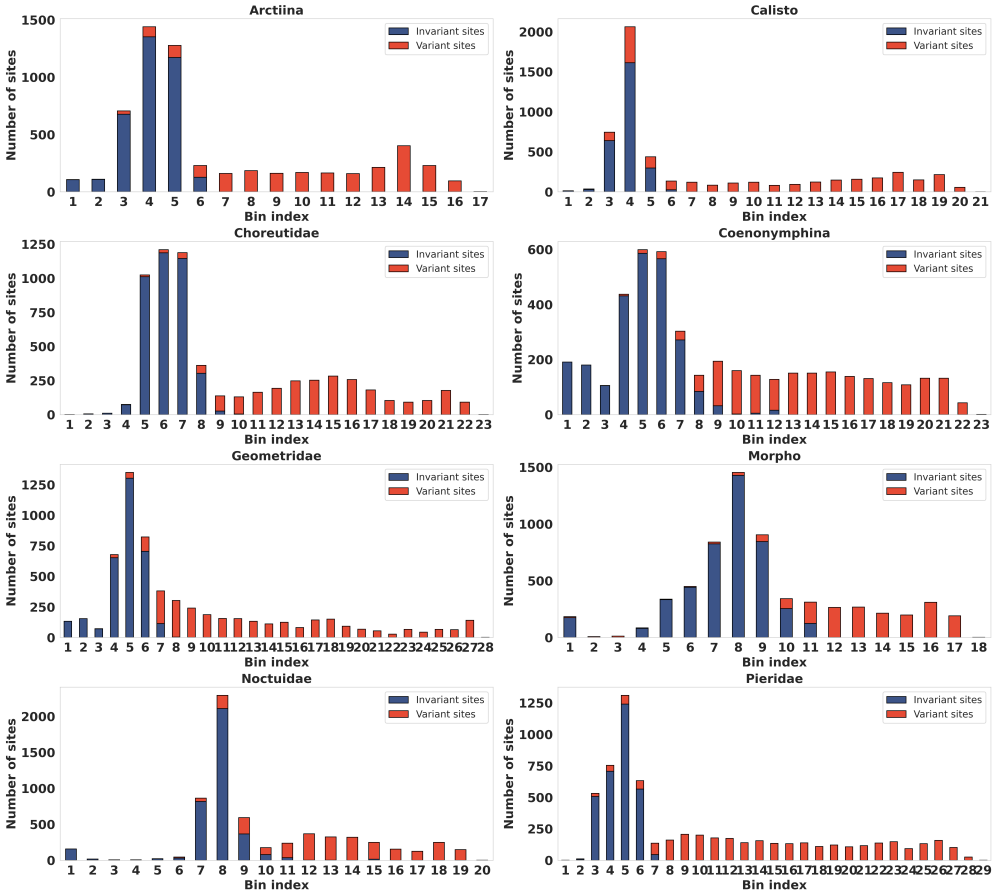


Fig. 2 The number of sites in each bin generated by PsiPartition on the different empirical DNA data sets (Arctiina, Calisto, Choreutidae, Coenonymphina, Geometridae, Morpho, Noctuidae, and Pieridae). Different bins have different number of sites, with a long tail of bins with small number of sites. Invariant and variant sites are in different colors. Note here a smaller bin index means a smaller PSI compared to the other bins, suggesting lower estimated evolutionary rates.

PsiPartition and PsiPartitionFast produce totally different partitioning compared to existing methods. The transfer distances between the partitions generated by our proposed methods and other existing methods are shown in Fig 3. Here the X-axis and Y-axis represent the partitions generated by different methods and blocks with smaller values and darker colors mean more similarities between the two partitions. The results demonstrate that the partitions generated by our proposed methods, PsiPartition and PsiPartitionFast, have quite large transfer distance to the partitions generated by remaining methods. On some data sets such as Morpho and Coenonymphina, the partitions generated by PsiPartition and PsiPartitionFast are also considerable different from each other. Interestingly, two existing methods, RatePartition-4 and

RatePartition-5 generated quite similar partitions to each other, because they utilized identical TIGER with only different division factors. Our PsiPartition instead optimized the PSI separately even on the same data set. These results suggest that our proposed methods are capable of searching for different partition schemes effectively and potentially finding the optimal ones.

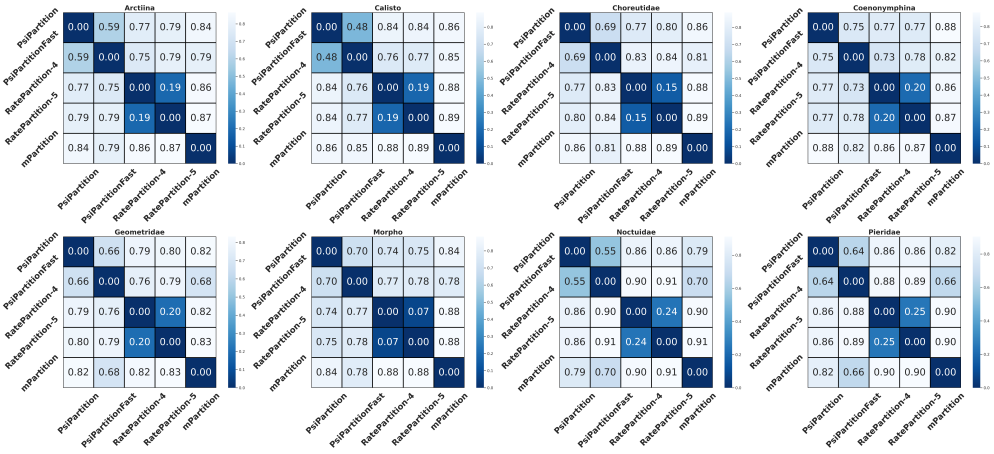


Fig. 3 The transfer distances between the partitions generated by our proposed methods (PsiPartition and PsiPartitionFast) and other methods on different empirical DNA data sets (Arctiina, Calisto, Choreutidae, Coenonymphina, Geometridae, Morpho, Noctuidae, and Pieridae). PsiPartition and PsiPartitionFast have large transfer distances to the other methods, suggesting their partitions are significantly different from the other methods.

Our proposed methods are also superior on the protein data. We evaluate the performance comparison of different partitioning methods on the empirical protein data sets, as shown in Fig 4. On most data sets, our proposed PsiPartition and PsiPartitionFast significantly reduce the BIC and AICc, as shown in Supplementary Table S2. PsiPartition significantly outperforms other methods in terms of the ΔBIC and ΔAICc. The only exception is on the data set "Ngyuen_2016e" where the PsiPartition is slightly worse than the RatePartition-4 and RatePartition-5. This is because the Ngyuen_2016e data set has a relatively small number of sites (385 amino acids) while others have about 600 sites or more. As the number of sites increases, the heterogeneity among sites could increase. The improvements achieved by our proposed methods compared to existing methods become more significant. Additionally, the PsiPartitionFast also achieved comparable performance to other methods. It is slightly worse than the PsiPartition because it does not adopt Bayesian optimization. The results demonstrate that our proposed PsiPartition is more accurate than the default IQ-TREE and other methods. By comparing PsiPartition and PsiPartitionFast, it is suggested that Bayesian optimization is necessary for accurate phylogenetic inference.

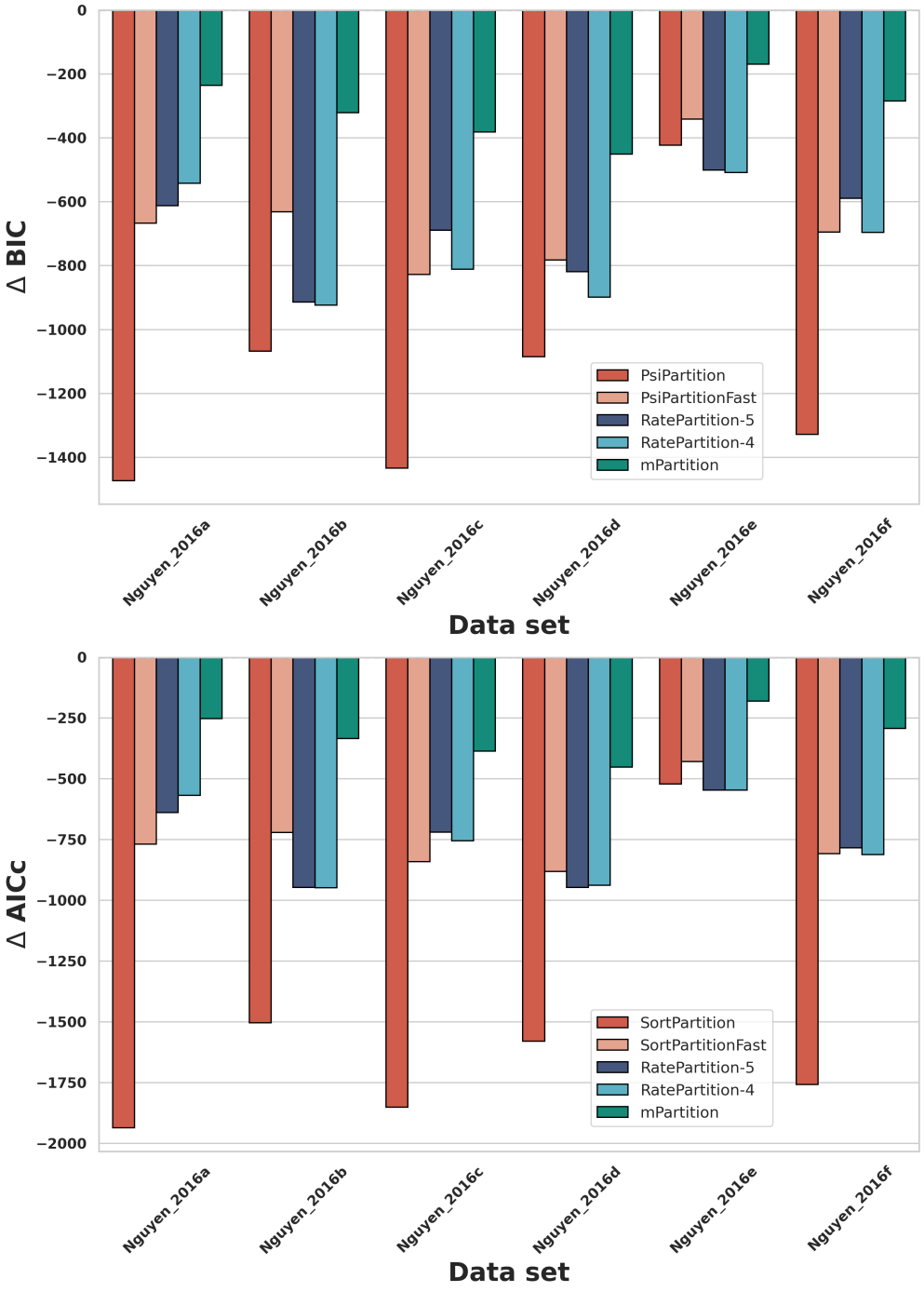


Fig. 4 The performance comparison of different partitioning methods on the empirical protein data sets (Ngyuen_2016[a-f]). Our proposed method PsiPartition achieves the best performance in terms of the ΔBIC and ΔAIC values, significantly outperforming the other methods.

As shown in Supplementary Fig. S1, the number of sites in each partition generated by PsiPartition on the protein data sets are also imbalanced. The invariant sites are distributed in different partitions and the first several partitions contain slow-evolving variant sites as well. Note here we apply our proposed PSI without noise term to estimate the site rates, suggesting that the noise term is not necessary for protein data sets.

3.2 PsiPartition accurately reconstructs simulated phylogenetic trees

We evaluate the performance of our algorithm on simulated data sets. We generate 100 alignments, each of which has 5,000 sites and 49 sequences. For each alignment, we randomly generate a tree topology as described in Supplementary Algorithm S2. The tree branch lengths are independently sampled from the exponential distribution with probability density $f(x) = e^{-x}$ for $x > 0$, as exponential distribution was suggested as the most fitted Venditti et al (2010). We employ Seq-Gen software Rambaut and Grass (1997) and the General Time Reversible (GTR) model Yang (2006) with 4-category Gamma distribution to generate the alignments. We utilize Seq-Gen's parameters "-r rate_matrix -f base_freq", where rate_matrix and base_freq are 6- and 4-dimensional vectors. Their elements are randomly generated from a uniform distribution $U[0, 1]$ and represent the rate matrix and base frequencies, respectively. For base_freq, an additional L_1 normalization is performed to make them indeed base frequencies.

The simulated data set is generated with heterogeneous evolutionary models. For each i from 1 to 10, we split 5,000 sites into i equal-length loci, and each locus is independently and randomly generated with Seq-Gen from the aforementioned process. This process is repeated 10 times for each i and finally, we obtained 100 simulated alignments. We denoted this data set as "SimDNA-1".

PsiPartition and PsiPartitionFast present a superior performance compared to other methods on the simulated data set SimDNA-1 in terms of the RF distance between the true trees and the reconstructed trees, as shown in Fig 5. The standard deviations of the RF distances are also shown. The results demonstrate that, for genomic data with heterogeneous site rates, our methods are robust and more accurate than the single homogeneous model without partitioning and existing partitioning methods. We emphasize that the differences of performance between our methods and others on the data sets with only one locus are not significant while the differences become larger in the data sets with more loci. Such differences do not increase when the number of loci continues to increase because the total lengths of sequences are always 5,000. This suggests the necessity of partitioning the sites by PsiPartition for large data sets. It is worth noting that both RatePartition-4 and RatePartition-5 failed to run on some of the simulated data sets with 87% (87/100) success rate because some sites are incorrectly partitioned and only one state is observed in some partitions thus IQ-TREE failed to start phylogenetic inference.

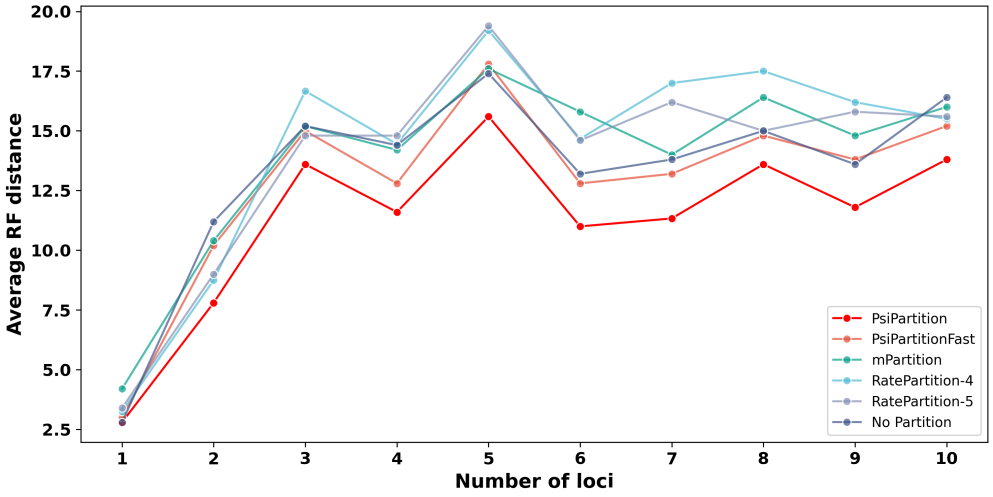


Fig. 5 The average RF distances between the true trees and the reconstructed trees on SimDNA-1. For each different number of loci, our proposed PsiPartition (red) reconstructs the most accurate trees, with the smallest RF distances. The advantages of PsiPartition compared to other methods are significant when the number of loci is larger than 2.

We also generate 100 simulated DNA alignments with 6 to 15 loci and the lengths of single locus are always set up to 1,000. That is, we obtain alignments that have a range of lengths from 6,000 to 15,000 and each length has 10 alignments. The algorithms and configurations to generate them are the same as aforementioned. We denote this data set as "SimDNA-2". We compare the performance of different methods in terms of RF distance, as shown in Fig 6. It is clear that our PsiPartition significantly outperforms other methods on most of the alignment. The only exception when the number of loci is 10 can be explained as a random bias during the phylogenetics inference because only 10 alignments are evaluated for each length. It is worth noting that some methods cannot reconstruct all phylogenies on SimDNA-2 (mPartition reconstructs 99 and RatePartition-5 reconstructs 92 out of 100 phylogenies).

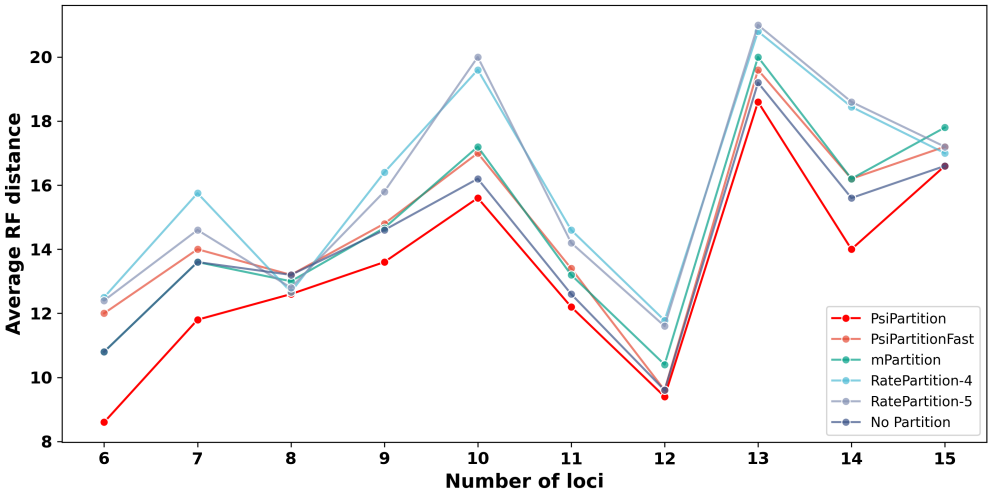


Fig. 6 The average RF distances between the true trees and the reconstructed trees on SimDNA-2. PsiPartition (red) reconstructs the most accurate trees with the smallest RF distances for all numbers of loci.

We further assess the performance of various methods on simulated DNA datasets (DataS2), as employed in previous studies Rota et al (2018); Le Kim and Le Sy (2020). These datasets were generated based on two types of trees: an asymmetrical tree—(outgroup, P, (O, (N, (M, (L, (K, (J, (I, (H, (G, (F, (E, (D, (C, (B, A)))))))))))))))—and a symmetrical tree—(outgroup, (((A, B), (C, D)), ((E, F), (G, H))), (((I, J), (K, L)), ((M, N), (O, P)))). In total, the datasets comprise 280 alignments, each consisting of 4,000 sites across 4 loci, with 1,000 sites per locus. The results are presented in Fig. 7. The reconstructed trees without partitioning (No Partition) exhibit the smallest RF distances to the true trees. Among the partitioning methods, mPartition demonstrates the highest accuracy. Our proposed methods, PsiPartition and PsiPartitionFast, follow closely, with accuracy nearly matching that of mPartition. RatePartition-4 and RatePartition-5 show the poorest performance among the partitioning methods. These findings indicate that both mPartition and our proposed methods outperform RatePartition-4 and RatePartition-5 in accuracy.

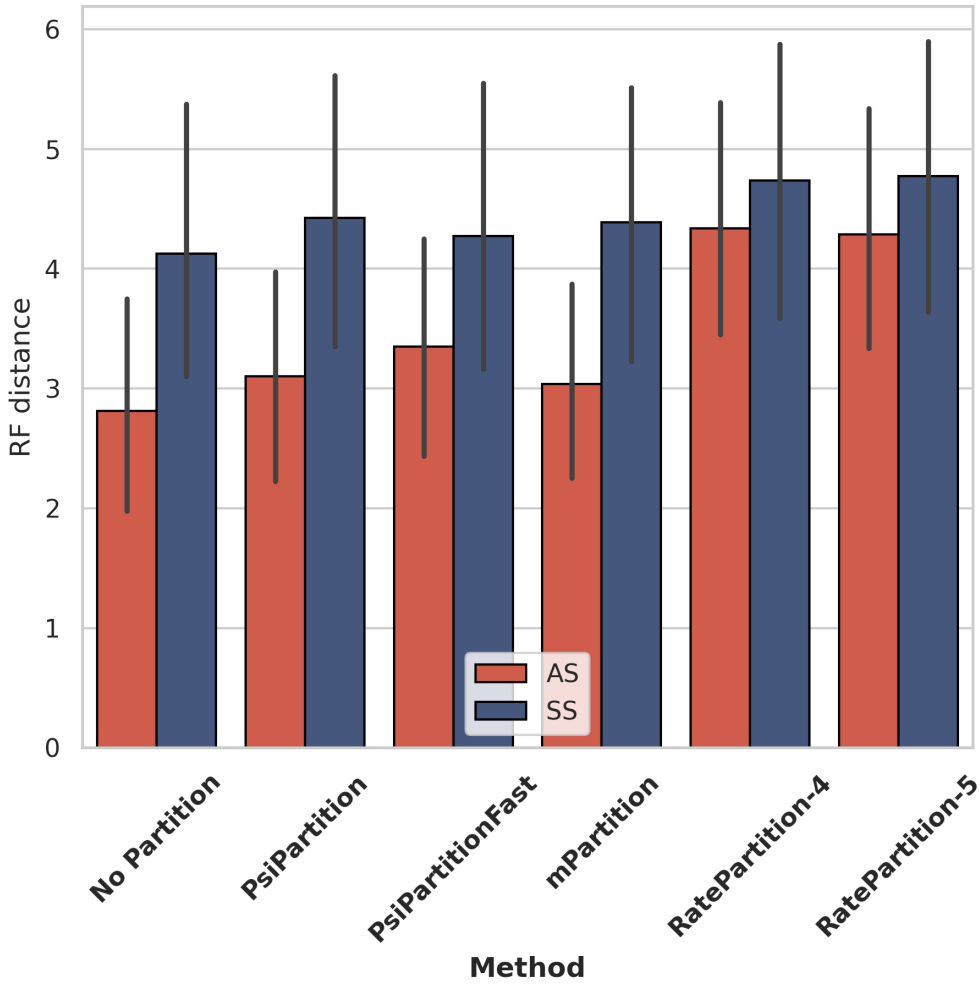


Fig. 7 The average RF distances between the true trees and the reconstructed trees on DataS2. AS denotes the asymmetrical tree and SS denotes the symmetrical tree.

3.3 Tree analysis on Noctuidae data set

We reproduced the phylogeny in the reported work Zahiri et al (2013) by IQ-TREE, and compared it to the phylogeny inferred by PsiPartition. Overall, the results show that the reconstructed tree based on PsiPartition has more likely topology and higher bootstrap supports of the branches, as shown in Supplementary Figure S3, compared to the tree without partitioning in Supplementary Figure S4. To emphasize the accuracy of PsiPartition clearly, we compare the reconstructed topologies and bootstrap supports in detail.

As shown in Fig 8, in the reproduced phylogeny (tree without partitioning), the species *Arcte modesta* and *Pseudoarcte melanis* are firstly categorized as a clade with a bootstrap support of 57 and then categorized with *Paracte schneideriana* into a clade with a bootstrap support of 100. The reproduced tree has the same topology and similar bootstrap support (65) to the tree in the reported work Zahiri et al (2013). Both of them are below the commonly used threshold of 70 and are not convinced. On the other hand, in the tree generated by PsiPartition, *Arcte modesta* and *seudoarcte melanis* are firstly categorized into a clade with a bootstrap support of 99 and then categorized with *Paracte schneideriana* into a clade with a bootstrap support of 100, which are more convincing, suggesting a possibly correct reconstruction of these species.

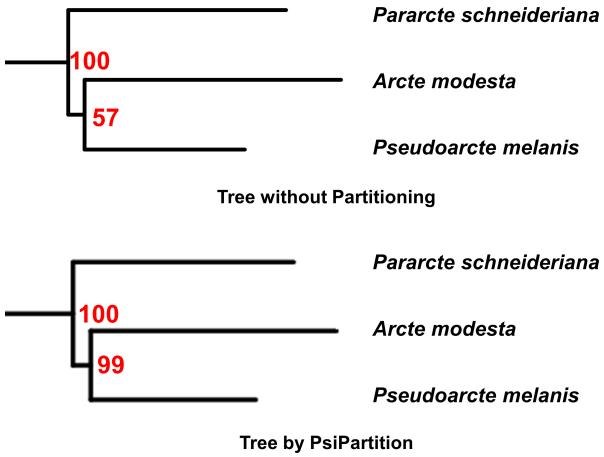


Fig. 8 The reconstructed trees of *Paracte schneideriana*, *Arcte modesta*, and *Pseudoarcte melanis* from phylogeny without partitioning (above) and PsiPartition (below), which shows higher bootstrap supports.

As shown in Fig 9, in the reproduced phylogeny (tree without partitioning), the species *Diloba caeruleocephala, Raphia abrupta or sp, Deltote uncula*, and *Cucullia unmbratica* are nested into three clades with bootstrap supports (43/81/59). In the reported work Zahiri et al (2013), *Deltote uncula* and *Cucullia unmbratica* are also categorized as a single clade with posterior probabilities of 0.61, which is very low. On the other hand, in the tree generated by PsiPartition, *Deltote uncula, Diloba caeruleocephala, Raphia abrupta or sp*, and *Cucullia umbratica* are nested into three clades with bootstrap supports (86/87/76). Overall, PsiPartition provides a different and possibly more accurate reconstruction of these species compared to the original work and reproduced work and reasons could be discussed in future work.

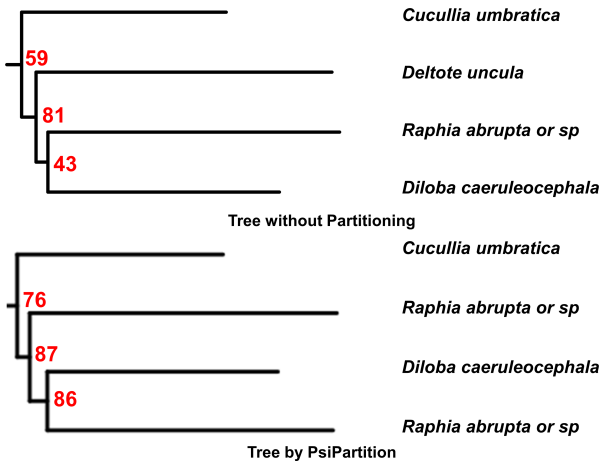


Fig. 9 The reconstructed clades of *Diloba caeruleocephala, Raphia abrupta or sp, Deltote uncula*, and *Cucullia unmbratica* from phylogeny without partitioning (above) and PsiPartition (below), which shows a possible correct topology with higher bootstrap supports.

As shown in Fig 10, both the reproduced phylogeny (tree without partitioning) and the tree generated by PsiPartition reconstructed the same topology for the species of *Craniophora ligustri, Polygrammate hebraeicum, Comachara cadburyi, Harrisimemna trisignata*, and *Cerma cerintha* yet with different bootstrap supports (98/100/62/51 v.s. 99/100/82/67). In the reported work Zahiri et al (2013), *Craniophora liguistr* together with *Harrisimemna trisignata* and *Cerma cerintha* are categorized into a clade with very low posterior probabilities of 0.59. PsiPartition thus provides an accurate reconstruction of this clade with higher bootstrap supports.

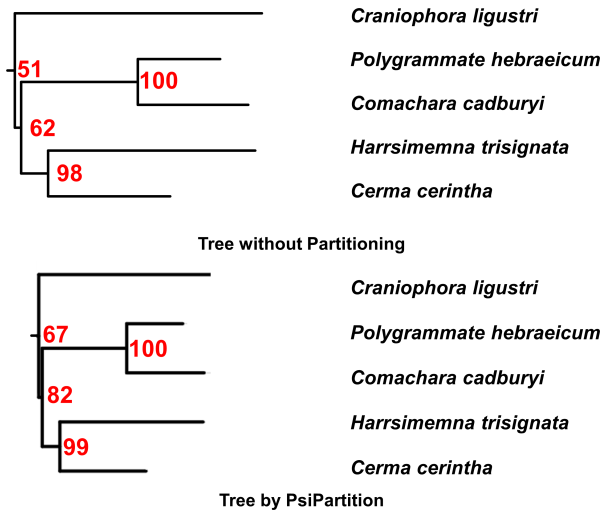


Fig. 10 The reconstructed clades of *Craniophora ligustri*, *Polygrammate hebraeicum*, *Comachara cadburyi*, *Harrisimemna trisignata*, and *Cerma cerintha* from phylogeny without partitioning (above) and PsiPartition (below). They have the same topology yet different bootstrap supports on some branches.

These results demonstrate that, by properly partitioning sites with our proposed PsiPartition method, we can improve the accuracy of the reconstructed phylogenetic trees and the bootstrap supports of the branches.

3.4 Efficiency of PsiPartition

We further assess the efficiency of PsiPartition on different empirical DNA data sets. The tracking of the evaluated BIC and AICc during the optimization iterations is shown in Fig 11. The results demonstrate that the Bayesian optimization gradually decreases BIC and finally converges (or fluctuates) at the optimal BIC values. AICc can be similarly analyzed. More specifically, we list the steps of iterations to achieve the best BIC and AICc in Table 3. These optimizations take a range of 18 to 30 steps (23.5 on average) to achieve an optimal BIC. Similarly, they take 18 to 29 steps (22.75 on average) to achieve an optimal AICc. The results illustrate that our proposed methods can expeditiously determine the optimal parameters in a limited steps of iterations.

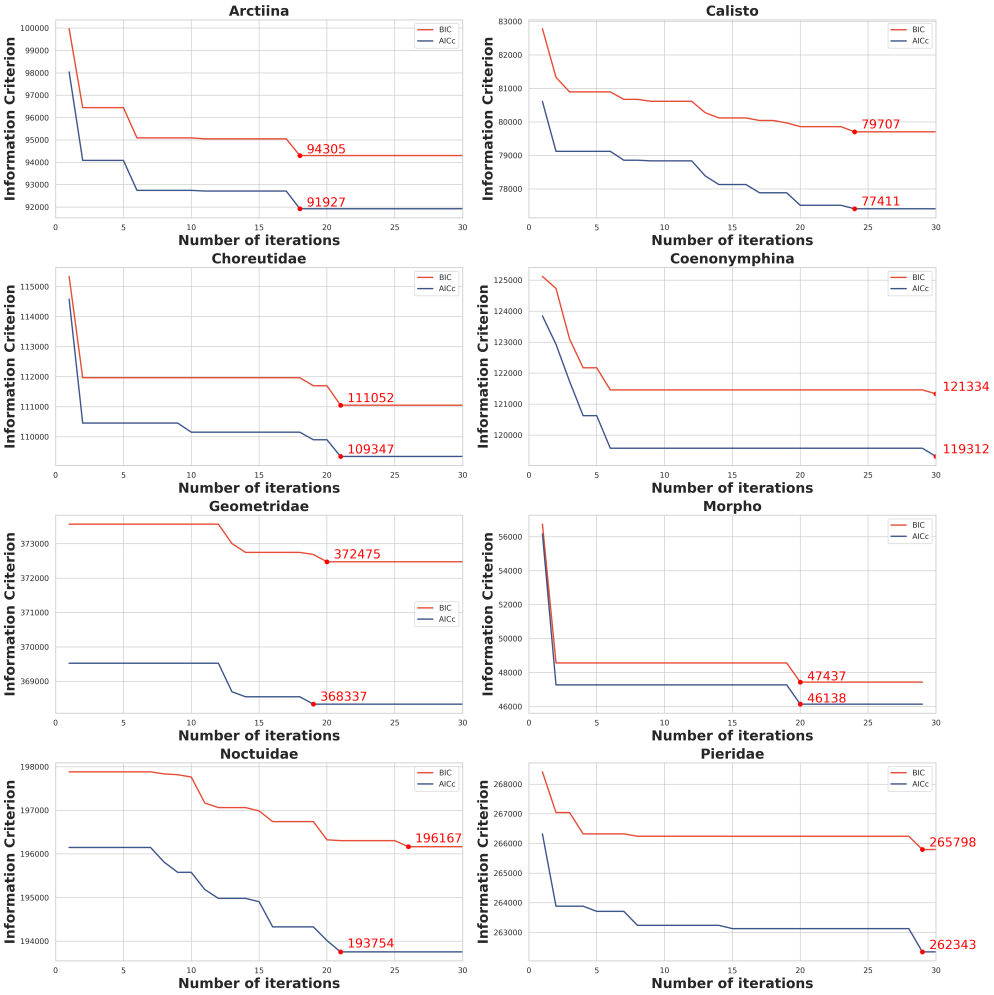


Fig. 11 Bayesian optimization for different empirical DNA data sets. In the curves, each point (k, v) where $k = 1, 2, \ldots, 30$ represent that PsiPartition achieved the minimum values of v of BIC (red) or AICc (red) in the first k iterations. The highlighted points with annotated values in the curves are the best information criterion values achieved by PsiPartition for the first time through the whole curves.

It is intractable to find the global minimum of BIC on all possible parameters, therefore we run PsiPartition 1000 times for the Morpho data set with uniformly sampled random parameters w from $[0, 1]^t$ and k from $\{2, 3, ..., k_{\max}\}$ where $t = 5$ and $k_{\max} = 30$, respectively. We check whether Bayesian optimization-based minimum BIC achieves the best among all randomly evaluated BICs. Considering the number of parameters of PsiPartition on DNA data sets is six, t-distributed Stochastic Neighbor Embedding (t-SNE) Van der Maaten and Hinton (2008) is employed to embed parameters into the 2-dimensional plane, which will keep the Euclidean distance between the

Table 3 The numbers of iterations to achieve the best BIC and AICc on different empirical DNA data sets.

Name	# Iter. to opt. BIC	# Iter. to opt. AICc
Arctiina	18	18
Calisto	24	24
Choreutidae	21	21
Coenonymphina	30	30
Geometridae	20	19
Morpho	20	20
Noctuidae	26	21
Pieridae	29	29

parameters as much as possible. We normalize each parameter to normal distribution before applying t-SNE to make them comparable. The BIC values of each evaluation are shown as the sizes and colors of the spots (darker and larger spots mean smaller BIC, and vice versa). As shown in Fig 12, it is clear that most of the evaluations from Bayesian optimization produce sufficiently good BIC values, which are concentrated into several clusters in the figure. The optimal parameters from Bayesian optimization (circled in orange) achieve the best BIC among all random evaluations. These results suggest that PsiPartition is capable of approaching the best parameters efficiently compared to the random search.

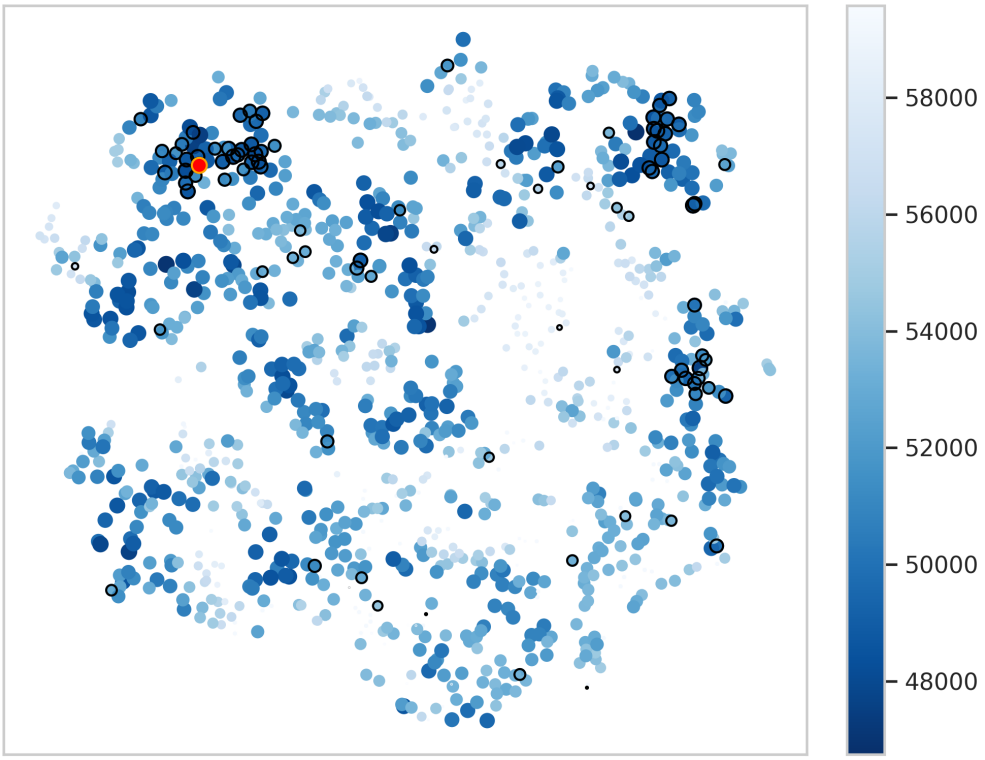


Fig. 12 Scatter diagram of BICs produced by PsiPartition with different parameters w and k on the Morpho data set. The trajectory of Bayesian optimization is circled with black and the best Bayesian optimized BIC is the red spot circled with orange.

We assessed the evolutionary time on an extensive alignment of Plethodontid, comprising 27 taxa and 15,788 base pairs gathered from 11 loci. RatePartition-4 and RatePartition-5 required 305.76 seconds (BIC 411794.14) and 311.05 seconds (BIC 412041.22), making them the most efficient methods among those evaluated. mPartition completed the analysis in 555.92 seconds, achieving a BIC of 416994.53. The running time per iteration is 196.55 seconds for PsiPartition and 62.00 seconds for PsiPartitionFast, which is similar to other methods (or shorter). However, PsiPartition and PsiPartitionFast took significantly longer for full optimization iterations, clocking in at 5896.61 seconds (BIC 396784.66) and 1859.91 seconds (BIC 410714.53), respectively. Since both PsiPartition and PsiPartitionFast rely on full phylogenetic inference of IQ-TREE, their processing time is proportional to the number of iterations. The time complexity of PSI is only a constant factor and remains manageable for medium-scale datasets. As an optimized parameter, a longer running time yields more accurate results for phylogenetic inference using PsiPartition and PsiPartitionFast.

4 Conclusion

In this study, we proposed a novel partition approach based on parameterized sorting indices (PSI) and Bayesian optimization to accurately and efficiently obtain the optimal partitioning of sites in genomic data. We applied our method to empirical data sets and achieved the best information criterion (BIC and AICc) compared to the models without partitioning or other existing state-of-the-art methods. We also show that Bayesian optimization-based PsiPartition is sufficiently efficient to determine the optimal parameters for partitioning the sites. In addition, we tested PsiPartition on the challenging simulated data sets with various site heterogeneity and demonstrated that our method outperforms other methods in terms of the average RF distance of the reconstructed phylogenetic trees to the actual trees. Finally, we show that on the empirical data set Noctuidae, our proposed method can improve the accuracy of the reconstructed phylogenetic trees, as well as the bootstrap supports of the branches. Overall, our proposed Bayesian optimization-based method achieves the best performance and provides a novel general framework to efficiently determine the optimal number of partitions, which can be easily extended to other partitioning methods. We hope its efficiency and effectiveness contribute to the growing demand for partitioning methods in phylogenetic analysis of large-scale genomic data.

Supplementary information. The online version contains supplementary material available at https://doi.org/xxx.

Acknowledgements. This work was supported by Hokkaido University DX Doctoral Fellowship (JST SPRING, Grant Number JPMJSP2119) to S.X. and SATREPS Project "Recovering High-Value Bioproducts for Sustainable Fisheries in Chile (ReBiS)" funded by JST/JICA (Grant Number JPMJSA2206) and JSPS KAKENHI Grant Number JP24H02213 in Transformative Research Areas (A) JP24A202 Integrated Science of Synthesis by Chemical Structure Reprogramming (SReP) and JP24K01533 to A.O.

Author Contributions. All authors conceived and designed the study. S.X. developed the method, performed data analysis, and wrote the paper. A.O. supervised the research project and critically reviewed and revised the manuscript. All authors approved the final manuscript.

Data availability. The corresponding reproducible source code and data sets are publicly available at https://github.com/xu-shi-jie/PsiPartition.

Declaration

Conflicts of interest. The authors declare no competing financial interest.

Ethical approval. This study does not involve any human or animal testing.

Consent for Publication. All authors read and approved the final manuscript.

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

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

Biewald L, et al (2020) Experiment tracking with weights and biases. URL https://www.wandb.com/, software available from wandb.com

Bofkin L, Goldman N (2007) Variation in evolutionary processes at different codon positions. Mol Biol Evol 24(2):513–521

Brandley MC, Schmitz A, Reeder TW (2005) Partitioned bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst Biol 54(3):373–390

Brown JM, Lemmon AR (2007) The importance of data partitioning and the utility of bayes factors in bayesian phylogenetics. Syst Biol 56(4):643–655

Chernomor O, Von Haeseler A, Minh BQ (2016) Terrace aware data structure for phylogenomic inference from supermatrices. Syst Biol 65(6):997–1008

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(6):833–844

Day WH (1981) The complexity of computing metric distances between partitions. Math Social Sci 1(3):269–287

Edger PP, Heidel-Fischer HM, Bekaert M, et al (2015) The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci 112(27):8362–8366

Felsenstein J (1981) Evolutionary trees from dna sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Frandsen PB, Calcott B, Mayer C, et al (2015) Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol Biol 15(1):1–17

Heikkilä M, Mutanen M, Kekkonen M, et al (2014) Morphology reinforces proposed molecular phylogenetic affinities: a revised classification for gelechioidea (lepidoptera). Cladistics 30(6):563–589

Ho SY, Lanfear R (2010) Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. Mitochondrial DNA 21(3-4):138–146

Höhna S, Landis MJ, Heath TA, et al (2016) Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol 65(4):726–736

Kaila L, Epstein ME, Heikkilä M, et al (2013) The assignment of prodidactidae to hyblaeoidea, with remarks on thyridoidea (lepidoptera). Zootaxa 3682(3):485–494

Kainer D, Lanfear R (2015) The effects of partitioning on phylogenetic inference. Mol Biol Evol 32(6):1611–1627

Kalyaanamoorthy S, Minh BQ, Wong TK, et al (2017) Modelfinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14(6):587–589

Kodandaramaiah U, Peña C, Braby MF, et al (2010) Phylogenetics of coenonymphina (nymphalidae: Satyrinae) and the problem of rooting rapid radiations. Mol Phylogenet Evol 54(2):386–394

Kristensen NP, Hilton DJ, Kallies A, et al (2015) A new extant family of primitive moths from k angaroo i sland, a ustralia, and its significance for understanding early l epidoptera evolution. Syst Entomol 40(1):5–16

Kumar S, Tao Q, Lamarca AP, et al (2023) Computational reproducibility of molecular phylogenies. Mol Biol Evol 40(7):msad165

Lanfear R, Calcott B, Ho SY, et al (2012) Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29(6):1695–1701

Lanfear R, Calcott B, Kainer D, et al (2014) Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol 14:1–14

Lanfear R, Frandsen PB, Wright AM, et al (2017) Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol 34(3):772–773

Le SQ, Gascuel O (2010) Accounting for solvent accessibility and secondary structure in protein phylogenetics is clearly beneficial. Syst Biol 59(3):277–287

Le Kim T, Le Sy V (2020) mpartition: a model-based method for partitioning alignments. J Mol Evol 88(8-9):641–652

Leavitt JR, Hiatt KD, Whiting MF, et al (2013) Searching for the optimal data partitioning strategy in mitochondrial phylogenomics: a phylogeny of acridoidea (insecta: Orthoptera: Caelifera) as a case study. Mol Phylogenet Evol 67(2):494–508

Letunic I, Bork P (2021) Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49(W1):W293–W296

Li C, Lu G, Ortí G (2008) Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci. Syst Biol 57(4):519–539

Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9(11)

MacQueen J, et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, pp 281–297

Matos-Maraví P, Núñez Águila R, Peña C, 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(1):1–18

Matsui M, Iwasaki W (2020) Graph splitting: a graph-based approach for superfamily-scale phylogenetic tree reconstruction. Syst Biol 69(2):265–279

McGuire JA, Witt CC, Altshuler DL, et al (2007) Phylogenetic systematics and biogeography of hummingbirds: Bayesian and maximum likelihood analyses of partitioned data and selection of an appropriate partitioning strategy. Syst Biol 56(5):837–856

Minh BQ, Nguyen MAT, Von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30(5):1188–1195

Nguyen AD, Gotelli NJ, Cahan SH (2016) The evolution of heat shock protein sequences, cis-regulatory elements, and expression profiles in the eusocial hymenoptera. BMC Evol Biol 16:1–13

Nguyen LT, Schmidt HA, Von Haeseler A, et al (2015) Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32(1):268–274

Ounap E, Viidalepp J, Truuverk A (2016) Phylogeny of the subfamily l arentiinae (l epidoptera: G eometridae): Integrating molecular data and traditional classifications. Syst Entomol 41(4):824–843

Penz CM, Devries PJ, 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(4):670–685

Poux C, Madsen O, Glos J, et al (2008) Molecular phylogeny and divergence times of malagasy tenrecs: influence of data partitioning and taxon sampling on dating analyses. BMC Evol Biol 8(1):1–16

Rajaei H, Greve C, Letsch H, et al (2015) Advances in g eometroidea phylogeny, with characterization of a new family based on p seudobiston pinratanai (lepidoptera, glossata). Zool Scr 44(4):418–436

Rambaut A, Grass NC (1997) Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. Bioinformatics 13(3):235–238

Rannala B, Yang Z (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol 43:304–311

Rasmussen CE, Williams CK, et al (2006) Gaussian processes for machine learning, vol 1. Springer

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53(1-2):131–147

Rönkä K, Mappes J, Kaila L, et al (2016) Putting parasemia in its phylogenetic place: a molecular analysis of the subtribe arctiina (lepidoptera). Syst Entomol 41(4):844–853

Rosenberg MS, Kumar S (2003) Taxon sampling, bioinformatics, and phylogenomics. Syst Biol 52(1):119

Rota J, Miller SE (2013) A new genus of metalmark moths (lepidoptera, choreutidae) with afrotropical and australasian distribution. ZooKeys (355):29

Rota J, Wahlberg N (2012) Exploration of data partitioning in an eight-gene data set: phylogeny of metalmark moths (lepidoptera, choreutidae). Zool Scr 41(5):536–546

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

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

Schwarz G (1978) Estimating the dimension of a model. The annals of statistics pp 461–464

Sihvonen P, Mutanen M, Kaila L, et al (2011) Comprehensive molecular sampling yields a robust phylogeny for geometrid moths (lepidoptera: Geometridae). PLoS One 6(6):e20356

Sugiura N (1978) Further analysis of the data by akaike's information criterion and the finite corrections: further analysis of the data by akaike's. Commun Stat- Theory Methods 7(1):13–26

Venditti C, Meade A, Pagel M (2010) Phylogenies reveal new interpretation of speciation and the red queen. Nature 463(7279):349–352

Wahlberg N, Wheat CW (2008) Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic dna extractions of lepidoptera. Syst Biol 57(2):231–242

Wahlberg N, Rota J, Braby MF, et al (2014) Revised systematics and higher classification of pierid butterflies (lepidoptera: Pieridae) based on molecular data. Zool Scr 43(6):641–650

Wandb (2024) https://github.com/wandb/sweeps/blob/master/src/sweeps/bayes_search.py, [Online; accessed 1-April-2024]

Williams C, Rasmussen C (1995) Gaussian processes for regression. Advances in neural information processing systems 8

Yang Z (2006) Computational molecular evolution. OUP Oxford

Zahiri R, Lafontaine D, Schmidt C, et al (2013) Relationships among the basal lineages of noctuidae (lepidoptera, noctuoidea) based on eight gene regions. Zool Scr 42(5):488–507