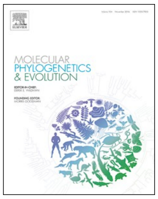
Molecular Phylogenetics and Evolution 154 (2021) 106966

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev



Evidence of absence treated as absence of evidence: The effects of variation in the number and distribution of gaps treated as missing data on the results of standard maximum likelihood analysis

Denis Jacob Machado[a,b,*], Santiago Castroviejo-Fisher[c], Taran Grant[d]

[a] University of North Carolina at Charlotte, College of Computing and Informatics, Department of Bioinformatcis and Genomics, 9201 University City Blvd, Charlotte, NC 28223, USA

[b] Universidade de São Paulo, Programa Interunidades de Pós-Graduação em Bioinformática, Rua do Matão, 1010, CEP: 05508-090 São Paulo, SP, Brazil

[c] Pontifícia Universidade Católica do Rio Grande do Sul, Laboratório de Sistemática de Vertebrados, Avenida Ipiranga, 6681, prédio 12, Partenon, CEP: 90619-900 Porto Alegre, RS, Brazil

[d] Universidade de São Paulo, Instituto de Biociências, Departamento de Zoologia, Laboratório de Anfíbios, Rua do Matão, tv. 14, 101, Cidade Universitária, CEP: 05508-090 São Paulo, SP, Brazil

ARTICLE INFO

Keywords:
Alignment
Ambiguous characters
Indel
Maximum likelihood
Missing data
Phylogenetic analysis

ABSTRACT

Although numerous studies have demonstrated the theoretical and empirical importance of treating gaps as insertion/deletion (indel) events in phylogenetic analyses, the standard approach to maximum likelihood (ML) analysis employed in the vast majority of empirical studies codes gaps as nucleotides of unknown identity ("missing data"). Therefore, it is imperative to understand the empirical consequences of different numbers and distributions of gaps treated as missing data. We evaluated the effects of variation in the number and distribution of gaps (i.e., no base, coded as IUPAC "." or "–") treated as missing data (i.e., any base, coded as "?" or IUPAC "N") in standard ML analysis. We obtained alignments with variable numbers and arrangements of gaps by aligning seven diverse empirical datasets under different gap opening costs using MAFFT. We selected the optimal substitution model for each alignment using the corrected Akaike Information Criterion in jModelTest2 and searched for optimal trees using GARLI. We also employed a Monte Carlo approach to randomly replace nucleotides with gaps (treated as missing data) in an empirical dataset to understand more precisely the effects of varying their number and distribution. To compare alignments, we developed four new indices and used several existing measures to quantify the number and distribution of gaps in all alignments. Our most important finding is that ML scores correlate negatively with gap opening costs and the amount of missing data. However, this negative relationship is not due to the increase in missing data per se—which increases ML scores—but instead to the effect of gaps on nucleotide homology. These variables also cause significant but largely unpredictable effects on tree topology.

1. Introduction

Standard maximum likelihood (ML) analysis of DNA sequences follows a three-step procedure composed of (I) multiple sequence alignment (MSA) using programs such as CLUSTAL X (Larkin et al., 2007), MAFFT (Katoh et al., 2005; Katoh and Toh, 2008), or MUSCLE (Edgar, 2004), (II) substitution model selection using programs like jModelTest (Posada, 2008) or PartitionFinder (Lanfear et al., 2012), and (III) tree searching using, for example, GARLI (Zwickl, 2006), PhyML (Guindon and Gascuel, 2003), RAxML (Stamatakis, 2006), or IQ-Tree (Nguyen et al., 2014). In the first step, insertion/deletion (indel) events are inferred according to user-specified indel opening and extension costs (GOC and GEC, respectively) and nucleotides inferred to be absent due to indels are represented in the alignment as gaps (coded as IUPAC "–"). In the second and third steps, gaps are treated as missing nucleotides and coded as ambiguities in the matrix (nucleotides of unknown identity; "?" or IUPAC "N"), thereby recasting evidence of absence as absence of evidence.

The effects of increasing amounts of ambiguity due to missing data are reasonably well understood: ML scores increase, the likelihood surface flattens, and, depending on the number and distribution of the ambiguities, topological relationships can change and support values

[*] Corresponding author at: University of North Carolina at Charlotte, College of Computing and Informatics, Department of Bioinformatcis and Genomics, 9201 University City Blvd, Charlotte, NC 28223, USA.

E-mail address: dmachado@uncc.edu (D. Jacob Machado).

https://doi.org/10.1016/j.ympev.2020.106966
Received 13 April 2020; Received in revised form 15 August 2020; Accepted 15 September 2020
Available online 22 September 2020

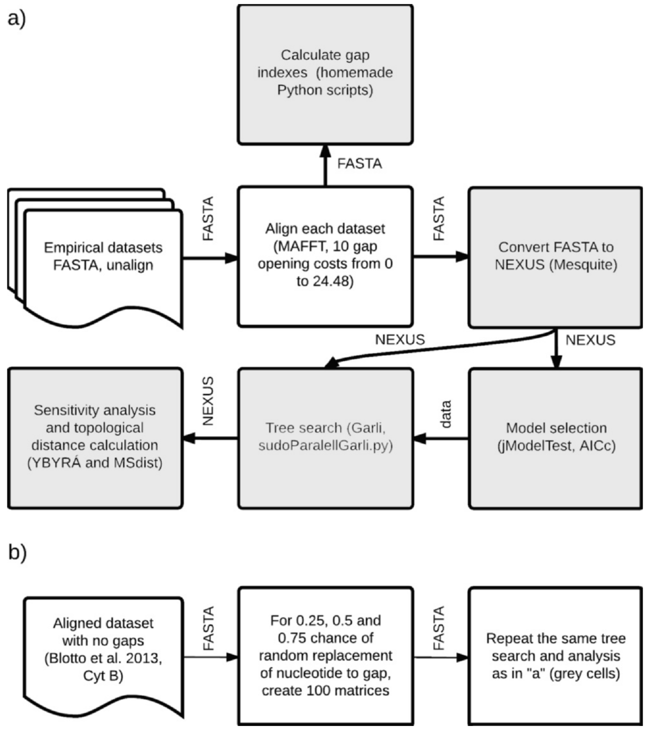

a)
Calculate gap indexes (homemade Python scripts)
FASTA
Empirical datasets FASTA, unalign
FASTA
Align each dataset (MAFFT, 10 gap opening costs from 0 to 24.48)
FASTA
Convert FASTA to NEXUS (Mesquite)
NEXUS
NEXUS
Sensitivity analysis and topological distance calculation (YBYRÁ and MSdist)
NEXUS
Tree search (Garli, sudoParalellGarli.py)
data
Model selection (jModelTest, AICc)
b)
Aligned dataset with no gaps (Blotto et al. 2013, Cyt B)
FASTA
For 0.25, 0.5 and 0.75 chance of random replacement of nucleotide to gap, create 100 matrices
FASTA
Repeat the same tree search and analysis as in "a" (grey cells)

Fig. 1. Summary of the methods for testing the effects of gaps treated as missing data in standard ML analysis. a) Empirical analysis with seven datasets (see Table 1). b) Analysis of 3,000 simulated matrices. Words by the arrows indicate the type of file or information being transferred. See text for additional details.

become inflated (Lemmon et al., 2009; Denton and Wheeler, 2012; Simmons and Norton, 2013; Simmons, 2014; Simmons and Goloboff, 2014; Sanderson et al., 2015). Similarly, previous studies have examined the degree to which different methods of alignment can alter results (e.g., Wheeler, 1994; Wheeler, 1995; Morrison and Ellis, 1997; Whiting et al., 2006; Wong et al., 2008; Blackburne and Whelan, 2012; Padial et al., 2014). However, to date no study has systematically investigated the effects of variation in the number and distribution of gaps treated as missing data on the results of standard ML analyses. Here, we evaluated the effects of this variation on model selection, ML score, and topology by analyzing highly variable alignments obtained by aligning diverse empirical datasets under different GOCs and by randomly replacing nucleotides with gaps.

2. Materials and methods

2.1. Datasets, model selection, and phylogenetic analysis

We summarized the analyses of the eight empirical datasets in Fig. 1 that varied extensively in the number of terminals and sequence length (Table 1). We aligned datasets 1–7 using the program MAFFT v7.147b. We performed the alignment of these seven datasets using the Needleman-Wunsch algorithm and 1,000 cycles of iterative refinement that incorporated global pairwise alignment information using the accuracy-oriented alignment method “G-INS-i.” This method is computationally intensive and therefore it is usually recommended for sequences of similar lengths with less than 200 terminals. This alignment strategy was implemented with MAFFT’s arguments “–globalpair” and “–maxiterate 1000”. We set the gap opening penalty to 0 and all values in a geometric progression of ratio 2 from 0.095625 to 24.48, including the program default value of 1.53 (all analyses employed a gap extension cost of 0.123). This resulted in 10 alignments per dataset that varied greatly in the number and distribution of gaps. Next, we converted the resulting MAFFT alignments from FASTA to NEXUS format using Mesquite v2.75 (build 564; Maddison and Maddison, 2015) and selected the optimal substitution model for each alignment using the corrected Akaike Information Criterion (AICc) in jModelTest v2.1.4. We then performed ML tree searches in GARLI v2.01 using default search parameters and assuming the optimal substitution model selected previously by jModelTest. Each tree search comprised 640 replicates spawned in parallel using a homemade Python script (sudoParallelGarli.py).

To clarify the effects of increasing amounts of ambiguity due to gaps and other possible alignment effects caused by the insertion of gaps during alignment, we performed additional Monte Carlo analyses using dataset 8 (Table 1). First, we aligned the data in MAFFT using the default GOC of 1.53 and the GEC of 0.123, trimmed the resulting alignment to include no leading or trailing gaps, and analyzed the matrix as described above. Next, we performed three rounds of 1,000 Monte Carlo replicates that randomly replaced nucleotides with gaps with probabilities of 0.25, 0.5, and 0.75, respectively, thereby increasing the number of gaps without altering nucleotide homology relationships or the length of the alignment (see below). Finally, we analyzed each of the resulting matrices in GARLI as described above, using the same substitution model selected for the original alignment.

All unaligned datasets and templates for the configuration files and execution scripts are available as supplementary material. We ran all compute-intensive analyses on the high performance computing cluster ACE, which is composed of 12 quad-socket AMD Opteron 6376 16-core 2.3-GHz CPU, 16 MB cache, 6.4 GT/s compute nodes (= 768 cores total), eight with 128 GB RAM DDR3 1600 MHz (16 × 8 GB), two with 256 GB (16 × 16 GB), and two with 512 GB (32 × 16 GB), and QDR 4x InfiniBand (32 GB/s) networking, housed at the Museu de Zoologia da Universidade de São Paulo.

2.2. Alignment characterization

We compared the 3,070 alignments generated for the eight datasets applying both the GOC used to generate the alignments and several new measures and indices that describe the number and distribution of gaps in each alignment. For each alignment, we calculated alignment length, total number of gaps, total number of characters (columns, positions, transformation series) containing gaps, mean number of gap openings, and number of identical characters per alignment (i.e., the number of columns that provide identical character-state distributions).

Table 1
Basic information of the eight datasets used in this study. * Effects of variation in the number and distribution of gaps treated as missing data; **Simulations.

Dataset	Reference	Marker	Length (bp)	# terminals	Analyses
1	Wheeler and Hayashi (1998)	18S rRNA	940–2020	32	*
2	Wheeler and Hayashi (1998)	28S rRNA	336–652	28	*
3	Healy et al. (2009)	18S rRNA	554–2189	58	*
4	Healy et al. (2009)	28S rRNA	668–4279	58	*
5	Wei et al. (2014)	12S mtDNA	437–1011	31	*
6	Mauro et al. (2014)	16S mtDNA	1534–1635	55	*
7	Pozzi et al. (2014)	mtDNA Control Region (CR)	515–2295	77	*
8	Blotto et al. (2013)	CytB	338–1003 (385)	88 (85)	**

Several indices and algorithms have been proposed to describe and compare multiple sequence alignments (e.g., Thompson et al., 2005; Kemena et al., 2011; Blackburne and Whelan, 2012; Soto and Becerra, 2014; Zambrano-Vega et al., 2017), some of which are derived from some of the same measures listed above. However, these methods focus on measuring overall genetic distances, structural modifications, or alignment accuracy or reliability, whereas we are specifically interested in evaluating the effects of varying the number and distribution of gaps. Consequently, we derived the following original indices (all defined to vary between 0–1) to summarize the distribution of gaps both within and among terminals.

Gap contiguity index (GCI).—GCI quantifies the degree to which gaps are grouped into contiguous strings or broken into short strings. For a given terminal with g gaps and g' trailing gaps (where trailing gaps are defined as all but the first in a contiguous string of gaps),

$$GCI = \frac{g'}{g-1}$$

GCI is undefined if the sequence has only one gap; otherwise, $GCI = 0$ if there is at least one gap or no gaps are contiguous (i.e., there are no trailing gaps), and $GCI = 1$ if there is $>$ all gaps are contiguous. For an alignment-wide value, we report the mean GCI for all terminals in the alignment. Sequences without gaps are ignored during GCI calculations. *Nucleotide contiguity index (NCI).*—NCI is equivalent to GCI but measures the contiguity of nucleotides instead of gaps. For a given terminal with n nucleotides and n' trailing nucleotides,

$$NCI = \frac{n'}{n-1}$$

NCI is undefined if the alignment is composed of $<$ 2 nucleotides; otherwise, $NCI = 0$ if all nucleotides for the terminal are separated by gaps and $NCI = 1$ if all nucleotides for the terminal are contiguous. For an alignment-wide value, we report the mean NCI for all terminals in the alignment.

Shared gaps index (SGI).—SGI quantifies the degree to which a given gap is shared among terminals. For a given character scored for t terminals of which t' terminals possess a given gap,

$$SGI = \frac{t'}{t-1}$$

Assuming the alignment is composed of at least two terminals and no columns consist entirely of gaps, $SGI = 0$ if the character contains no gaps and $SGI = 1$ if the gap is shared by all but one of the terminals (i.e., only one terminal possesses a nucleotide). For an alignment-wide value, we report the mean SGI for all characters that contain gaps (CG). *Topological gap index (TGI).*—TGI incorporates topological information that SGI omits. SGI summarizes the degree to which gaps are shared among terminals in the matrix. However, it ignores the topological distribution of those terminals—and, therefore, the topological distribution of the gaps—on the optimal tree. As such, for a given character with gaps shared by t' terminals and explained on the given tree by a minimum of t'' gap↔nucleotide transformations, TGI is defined as

$$TGI = \frac{t'}{t' \times t''}$$

TGI is undefined if the character does not contain gaps; otherwise, $TGI = 1$ if a minimum of one gap↔nucleotide transformation explains the gaps in all terminals (i.e., a single split divides all the terminals that possess the gap from all the terminals that possess a nucleotide) and decreases as the minimum number of gap↔nucleotide transformations increases. For an alignment-wide value, we report the mean TGI for all characters that contain gaps.

2.3. Evaluation criteria

We evaluated the effects of variation in the number and distribution of gaps treated as missing data in standard ML phylogenetic analysis by comparing the alignment parameters (i.e., GOCs) and measures and indices with three response variables: (I) optimal substitution model selected using jModelTest; (II) the optimal ML score from GARLI; and (III) the optimal tree topology. To assess the effect on tree topology, we calculated the match split distances (MSD) between the optimal topologies using MSdist v0.5 (Bogdanowicz and Giaro, 2012) and visualized their congruence using YBYRÁ (Machado, 2015). Note that GARLI collapses zero-length branches and, in these cases, MSdist will treat the nonbinary trees using the methodology described in Bogdanowicz and Giaro (2012) (2012: p. 158–159). We used R v3.3.1 (R Core Team, 2016) to fit linear models for correlation analysis.

3. Results

3.1. Alignments

We used the following GOC values: 0, 0.096, 0.191, 0.383, 0.765, 1.53 (1.53 is the program's default value), 3.06, 6.12, 12.24, and 24.48. We set the GEC a fixed value of 0.123. The different GOCs we used to align each dataset generated highly diverse alignments, as indicated by the variation in the values taken by all of the indices (Table S1). Among the alignments of sequences from datasets 1–7, mean GCI, mean NCI, mean SGI, and mean TGI values varied from 0.49–0.99, 0.38–0.99, 0.12–0.76, and 0.39–0.86, respectively. The most variable datasets for each of our indices were dataset 5 for mean SGI (0.39–0.58), dataset 6 for mean GCI (0.49–0.94), and dataset 7 for mean NCI (0.79–0.99) and mean TGI (0.394–0.571).

The 3,000 alignments generated by randomly replacing nucleotides with gaps in dataset 8 were also highly diverse. Alignment matrices composed of approximately 25% gaps had mean GCI, mean NCI, and mean SGI values of 0.23–0.27, 0.74–0.76, and 0.25–0.26, respectively. Alignment matrices with approximately 50% gaps had mean GCI, mean NCI, and mean SGI values of 0.49–0.51, 0.49–0.51, and 0.50–0.52, respectively. Lastly, alignment matrices with approximately 75% gaps had mean GCI, mean NCI, and mean SGI values of 0.74–0.76, 0.23–0.26, and 0.75–0.77, respectively. As expected, our results indicate that GCI and NCI are largely congruent with each other so that we can use any of them to predict the other.

3.2. Model selection

Despite the extensive variation among alignments, model selection varied little (Table S1). All models included gamma rate variation. Model selection chose the most complex model (GTR + I+G) for 80% of the alignments, including 100% of the alignments for datasets 3, 4, 6, and 7. Among the remaining datasets, we did not detect any trends in model selection. For example, dataset 5 varied most extensively, shifting between three models as GOCs increased: GTR + G for the two lowest gap opening costs, then HKY + G for the next three gap opening costs, GTR + G again for the next two, then GTR + I+G, GTR + G, and GTR + I+G for the three highest GOCs, respectively. In contrast, for dataset 2 the most complex model was chosen for the lowest two GOCs, then the less complex GTR + G, returning to the most complex model, then the even less complex HKY + G followed by GTR + G for the three highest GOCs.

3.3. Tree topology

Although variation in the number and distribution of gaps treated as missing data had little effect on model selection, it had a substantial effect on tree topology (Fig. 2). Nevertheless, we did not detect any pattern in the distribution of gaps to explain the observed variation in tree topology. Additionally, in many cases, the most distant topologies were derived from alignments with adjacent gap opening scores. Hence, we obtained significant differences in tree topology with only minor

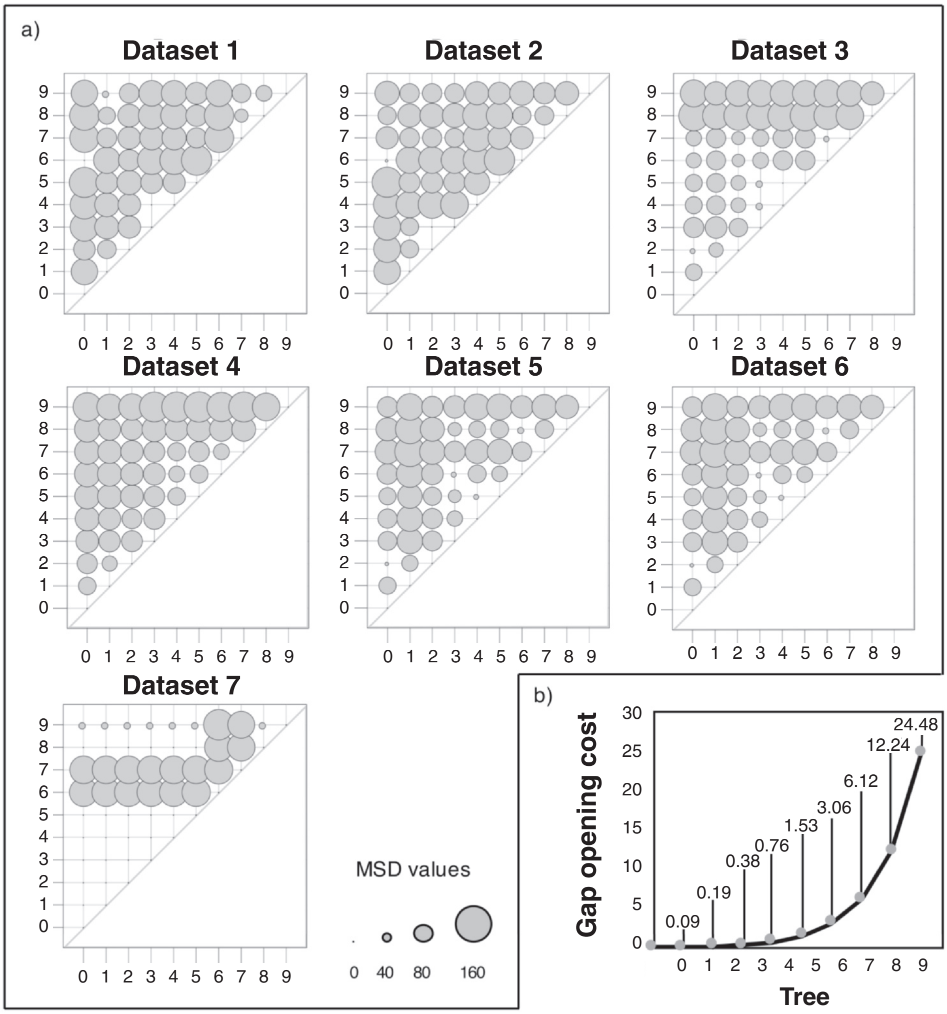


Fig. 2. Variation in tree topology according to sequence alignment in datasets with variable alignment length (1 to 7). a) Visualization of the topological distance between each pair of trees, organized in ascending order of the gap opening cost used for multiple sequence alignment (0 to 9). The larger the circle, the higher the average match-split distance (MSD); b) Gap opening costs used during multiple sequence alignment (0 to 24.48) and the corresponding tree number (0 to 9).

variations in the alignment parameters and the resulting number and distribution of gaps, even when there was no variation in the substitution model.

3.4. ML score

For most datasets, the ML score was negatively correlated with gap opening cost (adjusted $R^2 > 0.99$; Table S2). For all datasets except dataset 4, ML score was also negatively correlated with GCI (adjusted $R^2 = 0.73–0.99$) and NCI (adjusted $R^2 = 0.71–0.91$), both of which measure the degree to which sequences form contiguous strings or are broken into short strings. In contrast, the ML score positively correlated with alignment length, percentage of gaps, and mean SGI in all datasets except dataset 4 (Fig. 3).

Dataset 4 differed from all others in that the number of identical characters decreased as the gap opening cost increased (Fig. 4). The correlation analysis of the ML score and the mean TGI of dataset 4 had $R^2 = 0.98$. In contrast, the next-largest R^2 for this relationship was 0.86 for dataset 3 and the average R^2 for all datasets was 0.40 (see Table S2). In addition to that, the insertion of longer indels as the gap opening cost increases strongly affected nucleotide homology in dataset 4, leading to the unpredictability of mean GCI values. This also decreases similarity among characters in each alignment and results in alignments that differ more in the information contained in characters and their respective character states than in the distribution of gaps.

Although the average strength of the correlations between the aforementioned variables and the ML score was smaller than the correlation between gap opening cost and the ML score, we have no reason to assume the correlations are purely coincidental and instead propose that these variables partially account for the changes in the alignment matrix that lead to different ML scores. A special case seems to be when alignment is biased towards randomizing homology statements that follow long indels, as exemplified by dataset 4. In this case, the correlation of variables that explain the number and distribution of gaps in the alignment matrix with the ML score is weak, but we observed a strong correlation of the ML score and the TGI as a result of the number of gap↔nucleotide transformations on the tree.

3.5. Fixed nucleotide homology and alignment length

When we fixed the nucleotide homologies and alignment lengths, we observed a strong, positive, linear relationship between the number of gaps and the ML score. This means that ML score varied exclusively according to the number of indels treated as missing data, no matter the indel distribution patterns in the alignment. As such, there was no correlation between the ML score and any of the indices we defined, such as mean SGI (Fig. 5).

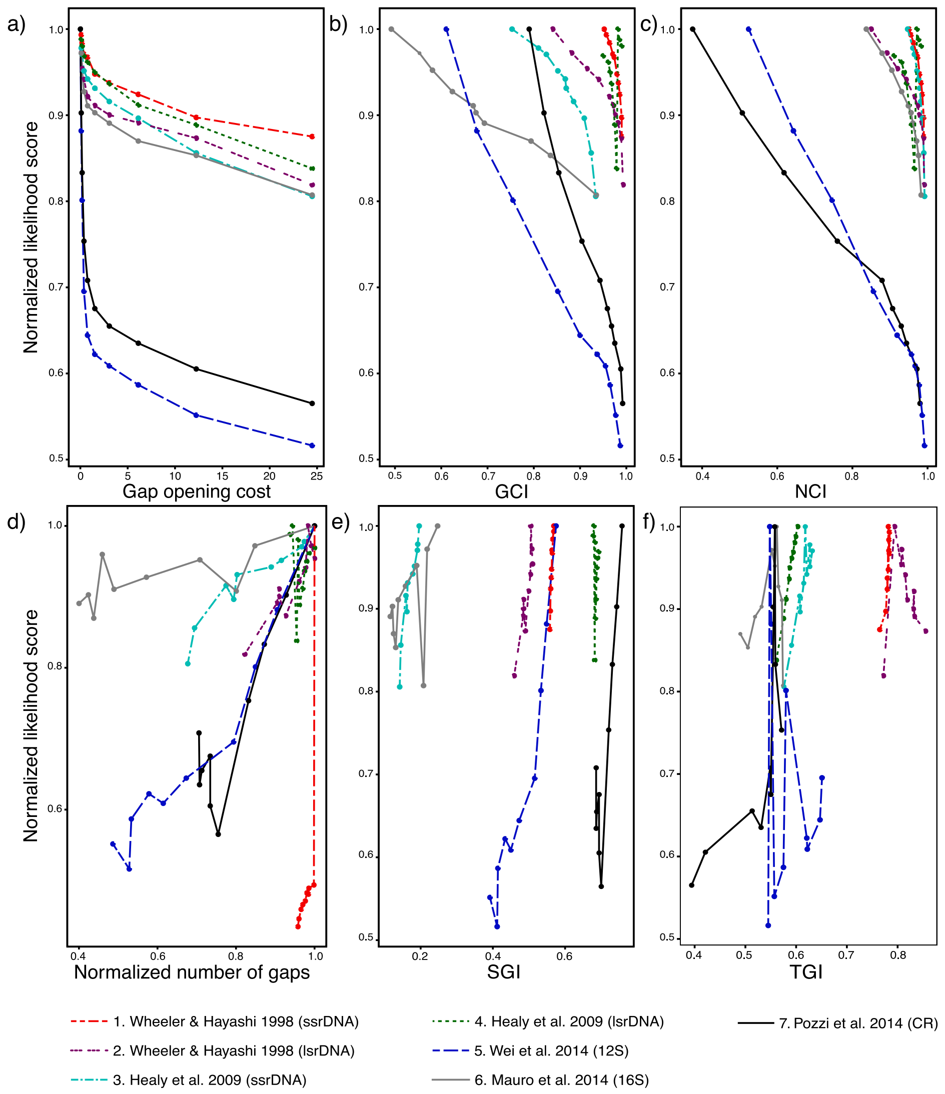


Fig. 3. Variation of different variables as a result of changes in the gap opening score and the likelihood score of the corresponding tree. The Y-axis shows the normalized likelihood scores. The X-axis shows a) the gap opening cost, b) the average gap contiguity index (mean GCI), c) the average nucleotide contiguity index (mean NCI), d) the normalized number of gaps (percentage), e) the average shared gap index (mean SGI), and f) the average topological gap index (mean TGI). The variable length was omitted since it closely resembles the variable percentage of gaps. Analyzed data and the results of the linear model analyses are available at Tables S1 and S2, respectively.

4. Discussion

This study is the first to systematically explore how the number and distribution of gaps treated as unknown nucleotides affect model selection, ML score, and topology in empirical phylogenetic analyses. Although the datasets we employed were small compared to many modern studies, they were chosen precisely because their relative simplicity facilitates interpretation, and our findings provide a basis for future studies to determine if the behavior of larger, more complex datasets is similar or different. Our general finding is that the effects depend on both the number of gaps and their effect on nucleotide↔nucleotide homologies. That is, all else being equal, as shown in our Monte Carlo simulations that randomly replaced nucleotides with gaps, increasing gaps results in higher ML scores and alignments approach trivial identity alignment (TIA; see Denton and Wheeler, 2012). However, in practice, introducing more gaps during alignment also affects the homology relationship among nucleotides, resulting in less predictable outcomes.

On the basis of our results, we identify three general responses to variation in the number and distribution of gaps. The first response, exemplified by analyses of datasets 1–3 and 5–7, occurs when sequence length is similar among all terminals and variation in the number and distribution of gaps has little effect on nucleotide↔nucleotide homologies. In this scenario, ML scores are negatively correlated with gap opening cost, number of gaps, sequence length, and mean GCI, positively correlated with mean SGI, and unrelated to mean TGI. The second response is observed in analyses of dataset 4, which has the greatest variation in sequence length (Table 1). In this response, ML score is negatively correlated with gap opening cost and positively correlated with mean TGI, which suggests that the insertion of long indels in these sequences strongly affected nucleotide homology. Finally, the third response is drawn from our Monte Carlo simulations, whereby we introduced gaps into the alignment matrix without altering nucleotide homology. In this case, ML scores improve as gaps increasingly replaced nucleotides, confirming that, all else being equal, ML score increases with the amount of missing data (cf. Denton and Wheeler, 2012).

In both responses 1 and 2, the uniformity of the nucleotide evolution models selected for the different alignments was unexpected. Given that alignments approaching TIA are simple matrices requiring few substitutions due to maximization of character columns that include only one nucleotide class (i.e., identical nucleotides and gaps treated as nucleotides of unknown identity), we expected that gappier alignments would require less complex models. Our interpretation of the lack of variation in model selection is that the alignments did not sufficiently approximate TIA to reduce the complexity of the models needed to explain the data. This explanation is supported by the fact that the most parameter-rich model was selected as optimal (i.e., GTR + I+G) for most alignments (75%).

We caution that our findings are agnostic with regards to the

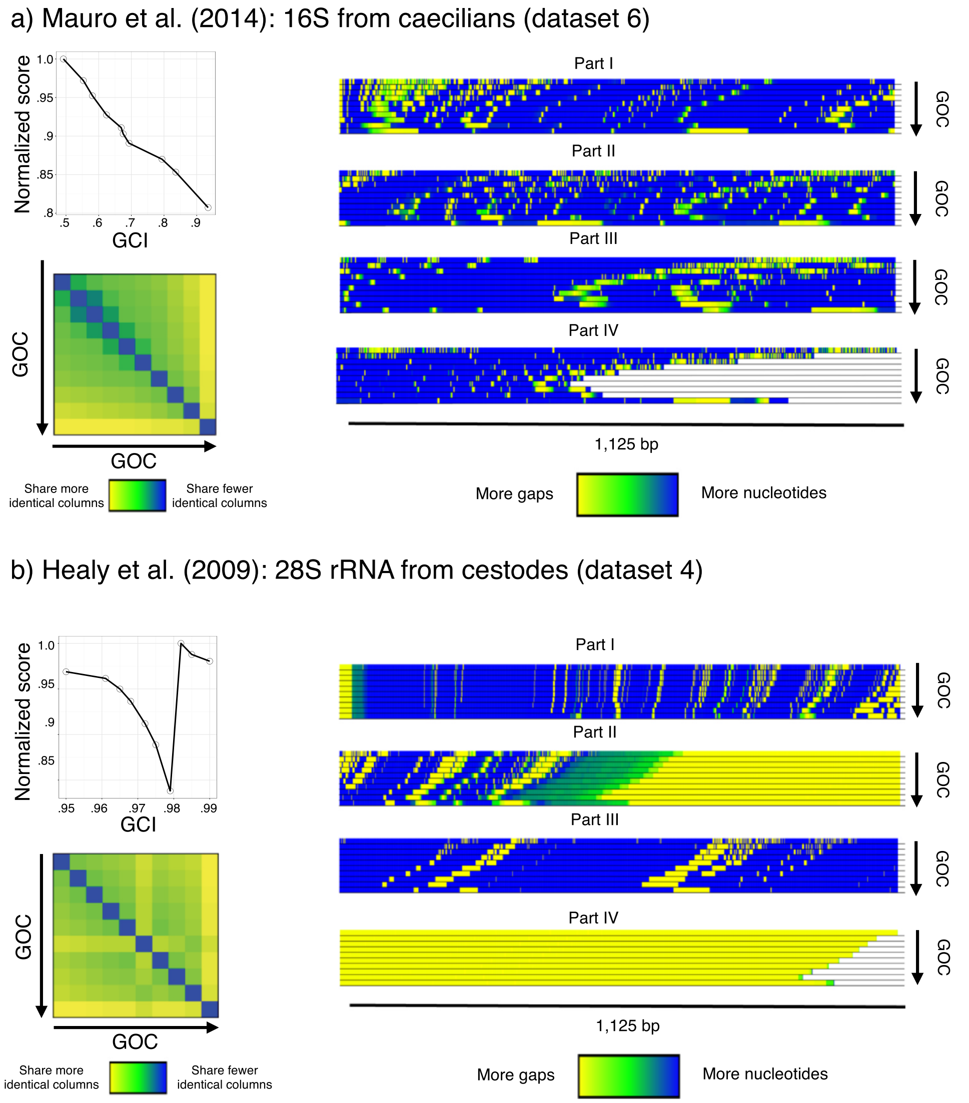


Fig. 4. As an example of all datasets except dataset 4, we show in a) the relationship of mean gap contiguity index (GCI) and the normalized likelihood scores (top left), the variation in the number of identical characters and the gap opening cost (GOC; see heatmap on the bottom left), and the distribution of gaps and nucleotides on all alignments (right) for datasets 6 (Mauro et al. 2014: 16S rRNA). Alignments are stacked on top of each other, ordered according to GOC, and divided into four windows of 1,125 bp. In b) we show the same information for dataset 4 (Healy et al. 2009: 28S rRNA), which differs from all others in the effect of gaps on nucleotide homology.

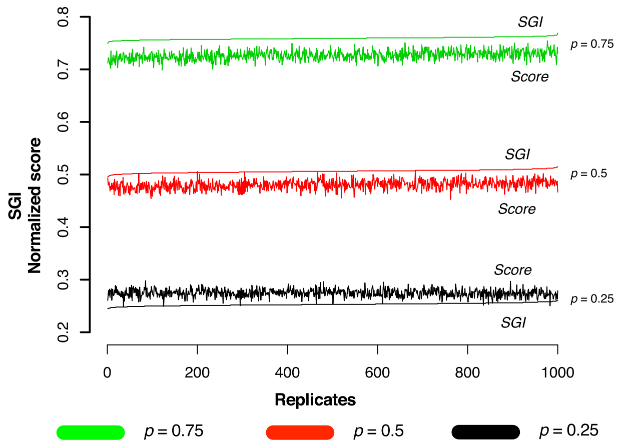


Fig. 5. Variation of normalized likelihood score (LS) and shared gaps index (SGI) of dataset 8 across three rounds of simulations (1,000 independent replicas each). In each simulation round, nucleotides were substituted by indels with a fixed probability (black = 0.25, red = 0.5, and green = 0.75). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

optimal gap opening and extension costs for empirical analyses that treat gaps as missing data. That is, although the effects of variation in gap costs on model selection, tree topology, and ML score can be predicted, none of these response variables provides a defensible optimality criterion for selecting alignments or alignment parameters in standard ML analysis. The program SATé (Liu et al., 2009; Liu et al., 2012) does employ ML score to choose among alignments obtained from MAFFT while treating gaps as unknown nucleotides, but Denton and Wheeler (2012) showed that the gaps-as-missing assumption results in TIA being optimal if alignments are evaluated on the basis of the ML score. In practice, it is highly unlikely for trivial alignments to be chosen as optimal in empirical studies because SATé searches using alignments obtained from MAFFT, which does not use ML as its optimality criterion and does not treat gaps as absence of evidence. Nevertheless, this does not absolve SATé of Denton and Wheeler's fundamental criticism, as its apparent immunity is due to its incomplete analysis of alignment space and inconsistent application of the optimality criterion. That is, given the specified optimality criterion, an adequately thorough analysis must select TIA as optimal, and it is only by employing different criteria for alignment and tree assessment that SATé avoids TIA. As Denton and Wheeler demonstrated, the problem is eliminated if gaps are attributed a cost in both the alignment and tree searching stages of analysis.

A long and growing list of theoretical and empirical studies has demonstrated the importance of treating gaps as indel events in phylogenetic analyses (e.g., Simmons et al., 2001; Ogden and Rosenberg, 2007; Dwivedi and Gadagkar, 2009; Dessimoz and Gil, 2010; Jordan and Goldman, 2012; Nagy et al., 2012; Yuri et al., 2013). Although this obviously does not prevent variation in the number and distribution of gaps from affecting results, by combining alignment and tree selection into a common analytical framework through generalized tree-alignment (Varón and Wheeler, 2013), gap opening and extension parameters can be chosen to maximize the likelihood score, as envisioned by Sankoff (1975) and implemented in programs like POY (Wheeler et al., 2015) and BEAST (Suchard et al., 2018). Nevertheless, the most common approach is to code gaps as unknown nucleotides. For example, all phylogenetic analyses of nucleotide sequences in the 50 open access articles published in Molecular Phylogenetics and Evolution since 2017 (available at www.journals.elsevier.com/molecular-phylogenetics-and-evolution/open-access-articles, accessed April 11, 2020) treated gaps as unknown nucleotides, as did all articles in Systematic Biology between 2013 and 2016.

Given how frequently gaps are treated as unknown nucleotides in phylogenetics, it is imperative to understand how their number and distribution affect results. Our findings are revealing, and there is no empirical or theoretical reason to believe they are unique to the datasets and optimality criterion we employed. Nevertheless, studies using larger and more diverse datasets and additional optimality criteria, especially Bayesian inference, must be undertaken to assess their generality and discover additional effects.

CRediT authorship contribution statement

Denis Jacob Machado: Funding acquisition, Writing - original draft, Writing - review & editing, Methodology, Formal analysis, Investigation, Resources, Software, Validation, Data curation, Visualization, Project administration. **Santiago Castroviejo-Fisher:** Conceptualization, Funding acquisition, Writing - original draft, Writing - review & editing. **Taran Grant:** Conceptualization, Funding acquisition, Writing - original draft, Writing - review & editing, Methodology, Formal analysis, Investigation, Resources.

Declaration of Competing Interest

The authors declared that there is no conflict of interest.

Acknowledgements

Funding was provided by the Fundação de Amparo à Pesquisa do Estado de São Paulo (Grant Nos.: 2012/10000-5, 2013/09598-8, 2015/18654-2, 2018/15425-0) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (Grant Nos.: 306823/2017-9 and 312744/2017-0). We thank the organizers of the joint 35th Annual Meeting of the Willi Hennig Society and XII Reunión Argentina de Cladística y Biogeografia (Museo Argentino de Ciencias Naturales "Bernardino Rivadavia", Buenos Aires, Argentina, October 5–8, 2016), where this research was first presented. The ideas presented in this study benefited from discussions with Darrel Frost, John Denton, Jose Padial, Pedro Peloso, and Ward Wheeler. We thank Mark Simmons, Pedro Ivo Simões, Ward Wheeler, and an anonymous reviewer for their criticisms of the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.5281/zenodo.3968770.

References

Blackburne, B.P., Whelan, S., 2012. Measuring the distance between multiple sequence alignments. Method. Biochem. Anal. 28, 495–502.

Blotto, B.L., Nunez, J.J., Basso, N.G., Ubeda, C.A., Wheeler, W.C., Faivovich, J., 2013. Phylogenetic relationships of a Patagonian frog radiation, the Alsodes + Eupsophus clade (Anura: Alsodidae), with comments on the supposed paraphyly of Eupsophus. Cladistics 29, 113–131.

Bogdanowicz, D., Giaro, K., 2012. Matching split distance for unrooted binary phylogenetic trees. IEEE-ACM T. Comput. Bi. 9, 150–160.

Denton, J.S.S., Wheeler, W.C., 2012. Indel information eliminates trivial sequence alignment in maximum likelihood phylogenetic analysis. Cladistics 28, 514–528.

Dessimoz, C., Gil, M., 2010. Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol. 11, R37.

Dwivedi, B., Gadagkar, S.R., 2009. Phylogenetic inference under varying proportions of indel-induced alignment gaps. BMC Evol. Biol. 9, 211.

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

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

Healy, C.J., Caira, J.N., Jensen, K., Webster, B.L., Littlewood, D.T.J., 2009. Proposal for a new tapeworm order, Rhinebothriidea. Int. J. Parasitol. 39, 497–511.

Jordan, G., Goldman, N., 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol. Biol. Evol. 29, 1125–1139.

Katoh, K., Kuma, K.i., Toh, H., Miyata, T., 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9, 286–298.

Kemena, C., Taly, J.F., Kleinjung, J., Notredame, C., 2011. STRIKE: evaluation of protein MSAs using a single 3D structure. Method. Biochem. Anal. 27, 3385–3391.

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

Larkin, M.A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al., 2007. Clustal W and Clustal X version 2.0. Method. Biochem. Anal. 23, 2947–2948.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biol. 58, 130–145.

Liu, K., Raghavan, S., Nelesen, S., Linder, C.R., Warnow, T., 2009. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. Science 324, 1561–1564.

Liu, K., Warnow, T.J., Holder, M.T., Nelesen, S.M., Yu, J., Stamatakis, A.P., Linder, C.R., 2012. SATé-II: very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. Systematic Biol. 61, 90.

Machado, D.J., 2015. YBYRÁ facilitates comparison of large phylogenetic trees. BMC Bioinformat. 16, 204.

Maddison, W.P., Maddison, D.R., 2015. Mesquite: a modular system for evolutionary analysis. Version 2 (75), 2011.

Mauro, D.S., Gower, D.J., Müller, H., Loader, S.P., Zardoya, R., Nussbaum, R.A., Wilkinson, M., 2014. Life-history evolution and mitogenomic phylogeny of caecilian amphibians. Mol. Phylogenet. Evol. 73, 177–189.

Morrison, D.A., Ellis, J.T., 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa. Mol. Biol. Evol. 14, 428–441.

Nagy, L.G., Kocsube, S., Csanádi, Z., Kovacs, G.M., Petkovits, T., Vágvölgyi, C., Papp, T., 2012. Re-mind the gap! Insertion–deletion data reveal neglected phylogenetic potential of the nuclear ribosomal internal transcribed spacer (ITS) of fungi. PLoS One 7, e49794.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32, 268–274.

Ogden, T.H., Rosenberg, M.S., 2007. Alignment and topological accuracy of the direct optimization approach via POY and traditional phylogenetics via ClustalW + PAUP*. Systematic Biol. 56, 182–193.

Padial, J.M., Grant, T., Frost, D.R., 2014. Molecular systematics of terraranas (Anura: Brachycephaloidea) with an assessment of the effects of alignment and optimality criteria. Zootaxa 3825, 1–132.

Posada, D., 2008. jModelTest: Phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.

Pozzi, L., Hodgson, J.A., Burrell, A.S., Sterner, K.N., Raaum, R.L., Disotell, T.R., 2014. Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. Mol. Phylogenet. Evol. 75, 165–183.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Sanderson, M.J., McMahon, M.M., Stamatakis, A., Zwickl, D.J., Steel, M., 2015. Impacts of terraces on phylogenetic inference. Systematic Biol. 64, 709–726.

Sankoff, D., 1975. Minimal mutation trees of sequences. Siam J. Appl. Math. 28, 35–42.

Simmons, M.P., 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. Mol. Phylogenet. Evol. 80, 267–280.

Simmons, M.P., Goloboff, P.A., 2014. Dubious resolution and support from published sparse supermatrices: the importance of thorough tree searches. Mol. Phylogenet. Evol. 78, 334–348.

Simmons, M.P., Norton, A.P., 2013. Quantification and relative severity of inflated branch-support values generated by alternative methods: an empirical example. Mol. Phylogenet. Evol. 67, 277–296.

Simmons, M.P., Ochoterena, H., Carr, T.G., 2001. Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. Syst. Biol. 50, 454–462. URL https://www.jstor.org/stable/3070935.

Soto, W., Becerra, D., 2014. A multi-objective evolutionary algorithm for improving multiple sequence alignments. In: Brazilian Symposium on Bioinformatics. Springer, pp. 73–82.

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

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., A, R., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus. Evolution 4, vey016.

Thompson, J.D., Koehl, P., Ripp, R., Poch, O., 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61, 127–136.

Varón, A., Wheeler, W.C., 2013. Local search for the generalized tree alignment problem. BMC Bioinformatics 14, 66.

Wei, S.J., Li, Q., van Achterberg, K., Chen, X.X., 2014. Two mitochondrial genomes from the families Bethylidae and Mutillidae: independent rearrangement of protein-coding genes and higher-level phylogeny of the Hymenoptera. Mol. Phylogenet. Evol. 77, 1–10.

Wheeler, W., 1994. Sources of ambiguity in nucleic acid sequence alignment. In: Molecular Ecology and Evolution: Approaches and Applications. Springer, pp. 323–352.

Wheeler, W.C., 1995. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst. Biol. 44, 321–331. https://doi.org/10.1093/sysbio/44.3.321.

Wheeler, W.C., Hayashi, C.Y., 1998. The phylogeny of the extant chelicerate orders. Cladistics 14, 173–192.

Wheeler, W.C., Lucaroni, N., Hong, L., Crowley, L.M., Varón, A., 2015. POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics 31, 189–196.

Whiting, A.S., Sites Jr, J.W., Pellegrino, K.C., Rodrigues, M.T., 2006. Comparing alignment methods for inferring the history of the new world lizard genus Mabuya (Squamata: Scincidae). Mol. Phylogenet. Evol. 38, 719–730.

Wong, K.M., Suchard, M.A., Huelsenbeck, J.P., 2008. Alignment uncertainty and genomic analysis. Science 319, 473–476.

Yuri, T., Kimball, R., Harshman, J., Bowie, R., Braun, M., Chojnowski, J., Han, K.L., Hackett, S., Huddleston, C., Moore, W., et al., 2013. Parsimony and model-based analyses of indels in avian nuclear genes reveal congruent and incongruent phylogenetic signals. Biology 2, 419–444.

Zambrano-Vega, C., Nebro, A.J., Durillo, J.J., García-Nieto, J., Aldana-Montes, J.F., 2017. Multiple sequence alignment with multiobjective metaheuristics: a comparative study. Int. J. Intell. Syst. 32, 843–861.

Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. thesis. The University of Texas at Austin.