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\textsuperscript{a,b,\textsuperscript{*}}, Santiago Castroviejo-Fisher\textsuperscript{c}, Taran Grant\textsuperscript{d}

\textsuperscript{a} University of North Carolina at Charlotte, College of Computing and Informatics, Department of Bioinformatics and Genomics, 9201 University City Blvd, Charlotte, NC 28223, USA
\textsuperscript{b} Universidade de São Paulo, Programa Interunidades de Pós-Graduação em Bioinformática, Rua do Matão, 1010, CEP: 05508-090 São Paulo, SP, Brazil
\textsuperscript{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
\textsuperscript{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

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
1055-7903/ © 2020 Elsevier Inc. All rights reserved.
Basic information of the eight datasets used in this study.

Table 1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Reference</th>
<th>Marker</th>
<th>Length (bp)</th>
<th># terminals</th>
<th>Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wheeler and Hayashi (1998)</td>
<td>18S rRNA</td>
<td>940–2020</td>
<td>32</td>
<td>*</td>
</tr>
<tr>
<td>2</td>
<td>Wheeler and Hayashi (1998)</td>
<td>28S rRNA</td>
<td>336–652</td>
<td>28</td>
<td>*</td>
</tr>
<tr>
<td>3</td>
<td>Healy et al. (2009)</td>
<td>18S rRNA</td>
<td>554–2189</td>
<td>58</td>
<td>*</td>
</tr>
<tr>
<td>4</td>
<td>Healy et al. (2009)</td>
<td>28S rRNA</td>
<td>668–4279</td>
<td>58</td>
<td>*</td>
</tr>
<tr>
<td>5</td>
<td>Wei et al. (2014)</td>
<td>mtDNA</td>
<td>437–1011</td>
<td>31</td>
<td>*</td>
</tr>
<tr>
<td>6</td>
<td>Mauro et al. (2014)</td>
<td>16S mtDNA</td>
<td>1534–1635</td>
<td>55</td>
<td>*</td>
</tr>
<tr>
<td>7</td>
<td>Pozzi et al. (2014)</td>
<td>mtDNA Control Region (CR)</td>
<td>515–2295</td>
<td>77</td>
<td>*</td>
</tr>
<tr>
<td>8</td>
<td>Blotto et al. (2013)</td>
<td>CytB</td>
<td>338–1003</td>
<td>88 (85)</td>
<td>**</td>
</tr>
</tbody>
</table>
Several indices and algorithms have been proposed to describe and compare multiple sequence alignments (e.g., Thompson et al., 2005; Kemen 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 contingency 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 no gap or 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 contingency 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

$$TGI = \frac{t'}{t' \times t^2}$$

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
variations in the alignment parameters and the resulting number and
distribution of gaps, even when there was no variation in the sub-
stitution 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 re-
spective 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 cor-
relation 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 corre-
lation 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).
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. 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.

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
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 pre-
pdicted, none of these response variables provides a defensible optim-
ality 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 op-
timality 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 optim-
ality 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

![Diagram](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.

![Diagram](Fig. 5) Variation of normalized likelihood score (LS) and shared gaps index (SGI) of dataset 8 across three rounds of simulations (1,000 independent replications 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.)
phylogenetic analyses (e.g., Simmons et al., 2001; Ogden and Rosenberg, 2007; Dviwedi and Gadağkar, 2009; Dessimo and Gil, 2010; Jordan and Goldman, 2012; Nagy et al., 2012; Yurí 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 criteria 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 Biogeografía (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


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


