COMPARACIÓN EMPÍRICA DE DOS MÉTODOS BIOGEOGRÁFICOS: OPTIMIZACIÓN DE ÁRBOLES BASADA EN PARSIMONIA Y ANÁLISIS DE SUBÁRBOLES LIBRES DE PARALOGÍA

ADRIANA MARCELA MORALES GUERRERO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGÍA
BUCARAMANGA
2011

COMPARACIÓN EMPÍRICA DE DOS MÉTODOS BIOGEOGRÁFICOS: OPTIMIZACIÓN DE ÁRBOLES BASADA EN PARSIMONIA Y ANÁLISIS DE SUBÁRBOLES LIBRES DE PARALOGÍA

ADRIANA MARCELA MORALES GUERRERO

Trabajo de Grado para optar al titulo de Bióloga

DIRECTOR
DANIEL RAFAEL MIRANDA ESQUIVEL

CODIRECTOR
MALTE C. EBACH

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGÍA
BUCARAMANGA
2011

CONTENTS

PÁG

INTRODUCTION 9

1. MATERIALS AND METHODS 13
1.1 SIMULATED DATA SETS.....13
1.1.1 Patristic Distances between instances of paralogy 15
1.2 REAL DATA SETS.....15
1.3 IMPLEMENTATION OF METHODS 17
1.4 COMPARISON OF IMPLEMENTATIONS 18
1.5 PARAMETER ESTIMATION (EMPIRICAL DATA SETS) 18

2. RESULTS 20
2.1 SIMULATED DATA SETS.....20
2.1.1 Patristic Distances between instances of paralogy 22
2.2 REAL DATA SETS.....22

3. DISCUSSION.....24
3.1 MISSING AREAS.....24
3.2 MASTs 26
3.3 PARALOGY 27

CONCLUSIONS.....28

REFERENCES 29

ANNEXES.....35

LIST OF ANNEXS

PÁG.

ANNEX A. TABLES 36
ANNEX B. FIGURES 40
ANNEX C. APPENDIX.....55

RESUMEN

TITULO: COMPARACIÓN EMPÍRICA DE DOS MÉTODOS BIOGEOGRÁFICOS: OPTIMIZACIÓN DE ÁRBOLES BASADA EN PARSIMONIA Y ANÁLISIS DE SUBÁRBOLES LIBRES DE PARALOGIA[*]

AUTOR: ADRIANA MARCELA MORALES GUERRERO[**]

PALABRAS CLAVES: metodología biogeográfica, optimizaron de árboles basada en parsimonia, análisis de subárboles libres de paralogia, TreeFitter, LisBeth.

DESCRIPCION
Este estudio presenta una comparación entre dos métodos biogeográficos; la optimizaron de árboles basada en parsimonia (PTF) y el análisis de subárboles libres de paralogia, a través de sus implementaciones TreeFitter y LisBeth respectivamente, para encontrar sus similaridades y diferencias. Nosotros construimos set de datos simulados y colectamos set de datos reales de previas publicaciones, implementamos ambos métodos y calculamos tanto el número de nodos iguales y compatibles compartidos entre ambos como los parámetros que influencian la símilaridad topológica. Con los resultados obtenidos observamos que ambos métodos encuentran patrones de relaciones de áreas muy diferentes. El único caso donde es siempre posible encontrar un patrón idéntico de relaciones entre ambos métodos es cuando el set de datos no tiene más de una instancia de ambigüedad. Los resultados también muestran que la diferencia entre ambos métodos no es dependiente de la cantidad de ambigüedad en el set de datos, sino mas bien de la posición de la instancia de ambigüedad en cada areagrama inicial/TAC, es decir la diferencia entre ambos métodos es debida a su disimilaridad teórica en el tratamiento de la ambigüedad mas que la acumulación de ruido en los set de datos. Le marco teórico de estos métodos es tan diferente que los resultados similares podrían interpretarse como una coincidencia. Esto refleja la 'crisis de identidad' que sufre actualmente la biogeografía, Como un programa de investigación que no comparte objetivos metodológicos consistentes y usualmente resulta en patrones de relaciones contradictorios.

[*] Trabajo de Grado
[**] Facultad de Ciencias, Escuela de Biología, Director: Daniel Rafael Miranda Esquivel, Codirector: Malte C. Ebach

ABSTRACT

TITTLE: PARSIMONY-BASED TREE FITTING AND PARALOGY-FREE SUBTREE ANALYSIS COMPARED[*]

AUTHOR: Adriana Marcela Morales Guerrero[**]

KEY WORDS: Biogeographical methodology, Event-based tree fitting, LisBeth, Paralogy free subtree analysis, Treefitter.

This study presents a comparison between Parsimony-based tree fitting and Paralogy-free subtree analysis through their implementations LisBeth and Treefitter, to find the similarities and differences between both methods. We used simulated and real data sets to perform a comparison of the number of equal and compatible nodes shared between both implementations. We identified the properties that influenced the similarity topological between both methods and observed that they find very different results. The only case where is always possible to find a identical pattern of relationship is when the data sets have no more than one instance of ambiguity. But, our results show that the difference between both methods is independent on the amount of ambiguity in the data sets. The value of maximum similarity index is dependent on the position of the instance of ambiguity in the each original areagram/TAC, that is to say, the difference between both implementations is due to their theoretical dissimilarity for treatment of each type of ambiguity rather than the accumulation of noise in a data set. The theoretical framework of these methods are so different that similar results are best described as coincidences. This reflects the ídentity crisis' suffered by the biogeography. As a research program, biogeography does not share consistent theoretical or methodological aims and usually give contrasting results.

[*] Work Degree

[**] Faculty of Science, School of Biology, Director: Daniel Rafael Miranda Esquivel, Co-Director: Malte C. Ebach

INTRODUCTION

Comparative biogeography uses phylogenies and taxic distributions in order to discover the area relationships (Ebach and Humphries, 2002; Ebach *et al*., 2003) and shares influences from various areas of science such as geography, ecology, geology and molecular systematics. Those engaged in these sciences consider themselves to be geographers, geologists, ecologist and molecular biologists first and 'biogeographers' second (Crisci, 2001; Morrone, 2009; Parenti and Ebach, 2009). Due to the large amount of influences and aims derived from these varying fields, a diversity of methods and implementations have been developing over the years (see Morrone, 2009). Based on this, biogeography could be divided into numerous subdivisions like paleobiogeography, ecography, event-based methods, cladistic biogeography and phylogeography. Therefore, there are many classification systems for such subdivisions (see Ronquist and Nylin, 1990; Morrone and Crisci, 1995; Crisci, 2001; Van Veller and Brooks, 2001), but maybe the largest grouping can be made between those methods that describe areas and discover their biotic relationships and, those methods that find the pattern of area relationships and the explanation to the mechanisms of distributions.

Comparative biogeography has a large diversity of methods and implementations, but results from these methods do not tend to converge on a single response, even conflicting results have been found. Therefore, is important to compare these methods to assess their similarities and differences. There are several comparisons of the biogeographic methods, (see Morrone and Carpenter, 1994; De Jong, 1998; Brooks and McLennan, 2001; Ebach and Edgecombe, 2001; McLennan and Brooks, 2002; Dowling et al., 2003; Villalobos, 2006; Sanmartín, 2007; Garzón-Orduña *et al*., 2008; Fattorini, 2008 and others), some of them dismiss the implementation of one method (i.e., Garzón-Orduña *et al*., 2008), while

other comparisons find no single methods to be consistently better (i.e., Morrone and Carpenter, 1994). Some recent biogeographical methods, namely Parsimony-based tree fitting (Ronquist, 2003a) and Paralogy-free subtrees analysis (Nelson and Ladiges, 1996) however, have not been compared. The aim of this paper is to compare both these methods through their implementations, namely Treefitter (Ronquist, 2003b) and LisBeth (Ducasse *et al*., 2008) to find the similarities and differences between them.

Rationale

A comparative biogeographical analysis comprises of three main steps:

1. Convert taxon cladograms into areagrams or taxon/area cladograms (TACs) by replacing the names of the taxa with areas in which they live
2. Compare areagrams or TACs
3. Derive the general areagram or General Area Cladogram (GAC)

Comparative biogeography is based on the assumption that each taxon is found in a single endemic area. Although correct, this condition is occasionally not met due to a large variety of problems such as poor species identification, inadequate area delimitation, conflicting information or ambiguity. While it is difficult to assess the quality of species identification and area delimitation, biogeographical methods however are able resolve the main sources of conflict and ambiguity. These are area duplications (i.e., geographical paralogy), Multiple Areas on a Terminal-branch, (MASTs or widespread taxa), and missing areas (i.e., areas absent from some of the original areagram or TAC) (Sanmartín and Ronquist, 2002; Ebach *et al*., 2005). Although these methods are able to deal with MASTs, duplication and missing areas, they treat this information very differently.

Parsimony-based Tree fitting (PTF)

Parsimony-based Tree fitting (PTF) is a biogeographical method, which seeks to find a pattern of historical distribution among multiples group of organisms, and to identify the causal mechanisms of this distribution over time (Sanmartín and Ronquist, 2002; Ronquist, 2003a; Sanmartín and Ronquist, 2004; Sanmartín, 2007; Sanmartín *et al*., 2007). PTF uses TACs namely, the phylogenetic trees which contain the distributions of taxa (Fig. 1a) (Sanmartín, 2007). TACs are compared to infer a common biogeographic pattern based on explicit distributional mechanisms. Two or more TACs combined form a GAC.

PTF involves explicit models of speciation, in which incorporates four types of events, vicariance (v), duplication (d), extinction (e) and dispersal (i). Each event is associated with a cost, and the GAC, will be the one that minimizes the total cost of the implied events (Sanmartín and Ronquist, 2002; Ronquist, 2003a; Sanmartín and Ronquist, 2004; Sanmartín, 2007). There are several four-event models that fit areas and phylogenetic tree of taxa together (i.e., Default 4E, Reconciliation, Maximum Codivergence and Fitch model). Ronquist (2003) showed that a general 4E model that maximizes cospeciation and duplication events, gives the best chances to find a pattern of area relationships under a wide range of conditions. That is because, these events are the only phylogenetically predictable mechanisms, since they are determinate by ancestor–descendant relationships within the phylogeny (Sanmartín *et al*., 2007). Treefitter program (Ronquist, 2003b) implement any event-costs models (Default 4E, Reconciliation, Maximum Codivergence and fitch models, etc.) but uses the Ronquist's model by default, (i.e., the default cost values are v = 0.01, d = 0.01, e = 1 and i = 2).

Paralogy-free Subtree Analysis (PSA)

Paralogy-free Subtree Analysis (PSA) is a biogeographical method developed by Nelson and Ladiges (1996) in order to remove geographic paralogy from areagrams. There are at least two implementations of this method, one TASS

software (Nelson and Ladiges, 1995), consists in removing geographic paralogy from areagrams, then uses Assumption 2 to resolve MASTs (i.e., Multiple Areas on Single Terminal-branches). TASS outputs the ambiguity-free subtrees as a matrix, which can be analysed using parsimony (e.g. TNT, NONA) or compatibility algorithms in order to obtain the GAC. More recently, PSA has been incorporated into the LisBeth program (Ducasse *et al*., 2008), which dispenses with Assumption 2 in favor of the Transparent Method (TM) to resolve MASTs (Ebach *et al*., 2005). Once the subtrees have been found, the program has the option to output a matrix representing the subtrees, this matrix could be analysed using parsimony or compatibility as well. The default option however, is to use three-item analysis, and compatibility coupled with an intersection tree to find a GAC.

The goal of PSA is to find area relationships without inferring biogeographical or evolutionary mechanisms *a priori*. PSA treats original areagrams as hypotheses of area relationships that contain no explicit information about the distribution and evolution of individual taxa (Fig. 1b) (Ebach and Humphries, 2002, Parenti and Ebach, 2009). When different original areagrams are compared, they overlap based on area relationships. Two or more overlapping areagrams form a general areagram (GAC), in which each component (i.e., junction between branches) represents a biotic divergence. Inferences as to the mechanisms that led to biotic divergence however are unobserved and therefore are unspecified (Ebach, 2003).

1. MATERIALS AND METHODS

We compared Paralogy-free Subtree Analysis (PSA) and Parsimony-based Tree fitting (PTF), through simulated and real data sets. We have used existing data sets from previous publications.

1.1 SIMULATED DATA SETS

To compare the performance of both implementations with different ambiguity levels, we constructed five hypothetical analyzes with different simulated data sets.

a. *A data set without ambiguous information* (Fig. 2): For this analysis we constructed a data sets with five areas and four original areagrams/TACs, each with five terminals, this data set was taken as negative control therefore we fitted each taxon to a single area.

b. *Data sets with one instance of ambiguity* (Fig. 3): For these analyzes, we evaluated three cases of ambiguous information, for each cases we constructed five data sets (Appendix 1). The data sets were different from the negative control as follows, for Case I (Paralogy only) we added one instance of paralogy in different positions to each data set. For including these instances of paralogy to the analyzes, we added one terminal at some initial areagram taken at random of each data set. For Cases II and III (MASTs only and Missing Areas only, respectively) we added one instance of MASTs and Missing Areas to each data set respectively, For including these instances of MASTs and Missing Areas to the analyzes, we deleted one terminal at some initial areagram taken at random of each data set.

c. *Data sets with more than one instance of ambiguity* (Fig. 4): For these analyzes, we evaluated six cases of ambiguous information, for each cases we constructed at least ten data sets (Appendix 2). the data sets were different from the negative control as follows, for Case I (paralogy only) we added two or more instances of paralogy in different positions to each data set (i.e., the number of instances of paralogy added were in a range from 3 to 8). For including these instances of paralogy to the analyzes, we added two or more terminals at some initial areagram(s) taken at random from each data set. For Case II (MASTs only) we added two or more instances of MASTs to each data set respectively (i.e., the number of instances of MASTs added were in a range from 2 to 8). For including these instances of MASTs in the analyzes, we eliminated two or more terminals from some initial areagram(s) taken at random or added one or more areas to each data set. For Case III (missing areas only) we added to each data set two or more instances of missing areas in different positions (i.e., the number of instances of missing areas added were in a range from 3 to 8). For including these instances of missing areas to the analyzes, we eliminated two or more terminals at some initial areagram taken at random or added one or more areas to each data set. For Case IV (Case with instances of paralogy + MASTs) we added two or more instances of paralogy and MASTs to each data set simultaneously (i.e., the number of instances of paralogy + MASTs added were in a range from 2 instances of paralogy and 2 instance of MASTs to 79 instances of paralogy and 79 instances of MASTs). For including these instances of paralogy + MASTs to the analyzes we increased the distribution of some taxa taken at random of each data set. For Case V (Case with instances of MASTs + missing areas) we added two or more instances of MASTs and missing areas in different positions to each data set simultaneously (i.e., the number of instances of MASTs + missing areas added were in a range from 1 instances of MASTs and 3 instances of missing areas to 7 instances of MASTs and 5 instances of missing areas). For including these instances of MASTs + missing areas to the analysis, we added one or more areas or eliminated two or more terminals from some initial areagram(s) taken at random to each data set. Finally

for the Case VI (Case with instances of paralogy + missing areas) we added to each data set two or more instances of paralogy and missing areas in different positions to each data set simultaneously (i.e., the number of instances of paralogy + missing areas added were in a range from 1 instances of paralogy and 1 instances of missing areas to 8 instances of paralogy and 8 instances of missing areas). For including these instances of MASTs + missing areas to the analysis we changed the distribution of some taxa taken at random to each data or added one area.

d. *A data set with the maximum of Paralogy and MASTs information* (Fig. 5): For this analysis we constructed a data sets with five areas and four original areagrams/TACs, each with five terminals, this data set was taken as positive control therefore we fitted each taxon to all areas.

1.1.1 Patristic Distances between instances of paralogy

We evaluated the influence of patristic distance of paralogy on the pattern of relationships obtained by both implementations. In these analyzes we constructed a simulated data set with 15 areas and two pectinate original areagrams/TACs each with 15 terminals. Next, two instances of paralogy to one original areagram/TAC selected at random were added, as shown in figure 6a. Finally patristic distance of the central paralogous area (A*) was increased with respect to its analogous areas, one component at a time.

1.2 REAL DATA SETS

We used seven additional real data sets from previous published studies. In each data set we transformed each phylogeny into an original areagram/TAC.

a) The Central American Poeciliidae fish genera *Xiphophorus* and *Heterandria* as used by Rosen (1978). Rosen's data represent the classical benchmark for biogeographical analysis as almost every method has been tested with them (Humphries and Parenti, 1999).
b) Six genera of Curculionidae distributed in the Subantarctic and Central Chilean subregions. Posadas and Morrone (2003) analyzed this data set using Brooks Parsimony Analysis (BPA), Dispersal-Vicariance Analysis (DIVA) and Reconciled Tree Analysis (RTA).
c) Seven phylogenies and distributions of Indo-Pacific coral reef biota used by Santini and Winterbottom (2002). The original data set was analyzed using BPA.
d) Eight groups of birds distributed in the Australian region (McLennan and Brooks, 2002). McLennan and Brooks (2002) analyzed this data set using primary and secondary BPA.
e) Phylogenies of nine Tibulidae groups distributed in western Mediterranean (de Jong, 1998). De Jong analyzed the original data set using BPA, CCA, Component Analysis (CA), Three-Area Statements (TTS) and PSA.
f) Thirty-seven phylogenies of taxa distributed in the Nearctic and Neotropical regions used by Escalante *et al.* (2007). Escalante et al., analyzed the original data set using PSA.
g) Fifty-four groups of extant non-marine taxa distributed in the Holartic region (Sanmartín *et al*., 2001). DIVA was originally applied to this data set

All phylogenies and distributions of each data set can be found in the original papers and the areas used correspond to previously defined areas of endemism by each author. All data sets satisfy the following criteria (Sanmartín *et al*., 2001):

(a) groups must be monophyletic
(b) At least three terminals of each original areagrams/TAC, must be occurring in the areas used in the analysis

(c) Original areagrams/TACs must comprise at least three terminals
(d) Original areagrams/TACs cannot have more than one polytomy or more than two trichotomies.

Besides the above some original areagrams in the data sets of Indo-Pacific, western Mediterranean, North Andean, Nearctic and Neotropical and Holartic were excluded from the analysis, due to the amount of MASTs in some of them, that produced many conflicting relationships. (see Parenti and Ebach 2009, p. 176).

1.3 IMPLEMENTATION OF METHODS

We analyzed the simulated and real data sets by using Treefitter version 1.2 (Ronquist, 2003b) and LisBeth (Ducasse *et al*., 2008). For LisBeth analyzes, we found the most compatible general areagram(s) (GA) through the minimizing incompatibilities algorithm, branch and bound technique and intersection of the resulting trees. On the other hand, with Treefitter, we used the default cost matrix (vicariance = 0.01, duplication = 0.01, extinction = 1 dispersion = 2) (Ronquist, 2003a) and the Recent option to treat widespread taxa. We found the best general area cladogram(s) (GAC) through a heuristic search, holding 100 trees in each step and using a neighbourhood of 30 nodes. Finally, when was necessary, we generated all possible dichotomous resolutions of polytomous TACs, each of them was weighted such that the sum of the alternatives corresponded to the value of the single fully resolved TACs (Sanmartín *et al*., 2001). We used the strict consensus tree when more than two taxon/area cladograms were produced.

1.4 COMPARISON OF IMPLEMENTATIONS

We compared the pattern obtained by LisBeth and Treefitter through the number of equal and compatible nodes shared between both implementations, equal nodes were then scaled in relation to the number of nodes obtained by each implementation, and the maximum value derived from the operation was considered the maximum similarity index (MS). A value 0 means that both implementations do not share patterns of area relationships, whereas a value 1 means that both pattern of relationships are equal, given the maximum number of possible shared nodes.

1.5 PARAMETER ESTIMATION (EMPIRICAL DATA SETS)

We estimated some properties only of the empirical data sets such as, number of original areagrams/TACs, number of terminals, number of areas and total instances of missing areas paralogy and MASTs. First we calculated the total instances of Paralogy and MASTs, for this, we estimated the maximum ambiguous information for each original areagrams/TACs in all data sets, as follows:

- Maximum instances of Paralogy for a original areagram/TAC = Number of taxa -1 multiplied by number of Areas
- Maximum instances of MASTs for a original areagram/TAC = Number of taxa multiplied by number of Areas -1

Then we calculated the instances of paralogy and MASTs observed for each original areagrams/TACs in all data sets, as illustrated in Fig. 7a, b. This values (Maximum instance of ambiguity and ambiguity observed) were added up independently and then we scaled the total ambiguity observed with respect to the

total maximum ambiguity and thus was obtained the total instances of paralogy and MASTs to the each data set (Table 1).

In the same way, we calculated the total instances of missing areas, first we estimated the minimum of missing information to each data set (Fig. 8a), as the number of areas multiplied by number of areagrams, then we calculated the instances of missing information observed as illustrated in Fig. 8b. This value was then scaled with respect to the maximum ambiguity, and thus was obtained the total of missing information to the each data set (Fig. 8b).

Finally we identified the properties that influenced the similarity topological through a multiple linear regression analysis. We used the selection criteria
R^2 and adjusted R^2

2. RESULTS

2.1 SIMULATED DATA SETS

For the first analysis without ambiguity we found that both analyzes obtained the same results, therefore we corroborated the efficiency of both implementations to apply the basic assumption of the comparative biogeography, “one-area, one-taxon”.

In the second analysis with one ambiguity instance, in each case (with their respective five data sets) both implementations obtained the same pattern of area relationships, but for the third analysis, with more than one instance of ambiguity, MS value behaved as follows:

- In Case I, (Paralogy only), we found that in three out of ten data sets, the patterns of area relationships obtained by both implementations were different, with a MS value equal to zero for each (Average:0.7, Median:0.5, Mode:1)
- In Case II, (MASTs only) ,we found that in five out of the ten data sets, the patterns of area relationships obtained by both implementations were different, the MS value was in a range of 1 to 0.33 (Average:0.68, Median:0.5, Mode:1)
- In Case III, (missing areas only), we found that in four out of ten data sets, the patterns of area relationships obtained by both implementations were different, with a MS value equal to zero for each one (Average:0.6, Median:0.5, Mode:1)
- In Case IV (Case with instances of Paralogy + MASTs) We found that the patterns of area relationships obtained by both implementations were the same (MS = 1) until the data set had 10 instances of Paralogy and 10 instances of MASTs, at this point the MS value started to diminish in a range from 0.66 to 0. When we added 44 instances of paralogy and 44 instances of MASTs,

Treefitter found a basal polytomy pattern, whereas LisBeth even found a pattern without polytomies. (Average:0.66, Median:0.66, Mode:1)

- In Case V (Case with instances of MASTs + “missing areas”) we found that in six of the ten data sets, the patterns of area relationships obtained by both implementations were different, the MS value was in a range of 1 to 0.25. (Average:0.66, Median:0.58, Mode:0.5)
- In Case VI (Case with instances of Paralogy + “missing areas”) we found that in three of the ten data sets, the patterns of area relationships obtained by both implementations were different, the MS value was in a range of 1 to 0.25. (Average:0.84, Median:0.58, Mode:1)

In this analysis we observed that the behavior of the maximum similarity index does not depended on the amount of ambiguity instances in the data sets. For example, in the case with paralogy only, when we constructed a data set with seven instances of ambiguity (Appendix 2, data set 5), we did not find different patterns of relationships between both implementations, but when we added five instances of paralogy to another data set (Appendix 2, data set 10) the patterns obtained by both implementations were different (MS= 0). Therefore these analyzes showed that there is not a pattern to indicate that the greater the number of ambiguity information, the lower the similarity between both implementations. On the other hand, it is important to note that the cases with paralogy were those that had the highest average value of MS (paralogy only: 0.7 and paralogy + missing :0.84), therefore there is a greater chance to finding similar results between both implementations if the data sets contains only instances of paralogy. With the other cases of ambiguity the implementations tend to present lower values of MS.

Finally in the fourth analysis, with the maximum observable ambiguity, both implementations found 105 general areagrams/GACs, the strict consensus tree (as applied to the trees found by Treefitter) and the intersection tree (as found by

LisBeth) was the same basal polytomy. In this case, none of both implementations found evidence for resolve the area relationships.

2.1.1 Patristic Distances between instances of paralogy

In this analysis we found that both analyzes yield, from identical patterns of relationships to completely different patterns of relationships. When we increased patristic distance of the central paralogous area (A*) to ≤ 3 with respect to their analogous areas, the pattern of area relationships were the same for both implementations, in this case Treeefitter found the events of duplication + sorting to resolve the instances of paralogy, and LisBeth removed this ambiguity. But when the patristic distance of the central paralogous area (A*) was greater than three with respect to their analogous areas, the MS value decreased (0.06). In this case whereas LisBeth presents a pattern equal to the initial (i.e., that is, the same pattern of relationships, that when the patristic distance increases to ≤ 3), Treefitter moves the paralogous area towards the root of the general areagram/GAC (Fig. 6b,c) and found the vicariance+dispersal events to explain the pattern.

2.2 REAL DATA SETS

In the most data sets, the values of MS index found between both implementations (LisBeth and Treefitter), was fewer than 0.5 (Average: 0.288, Median:0.375). Only in one case, in the data set of Central america (Rosen's data), the area relationships are identical with a value of 1, this case corresponds to one of the three biogeographical hypotheses found by Treefitter for this data set, compared with the intersection tree found by LisBeth. In the other data sets the MS values vary considerably. The properties and the maximum similarity found between both implementations for each data set is shows in table 2.

The parameters that influenced the maximum similarity between both implementations were, the number of areas, instances of paralogy, instances of MASTs and instance of missing areas, we found these four parameters by the two selection criteria in a multiple linear regression (Table 3). These results are consistent with the results derived from simulated data set because, even when the instances of ambiguity and the number of areas were the properties that influences the similarity, there is no direct correlation between these properties and the MS (Table 2). Rather, there is a group of variables that together and in different ways influences the similarity between the two implementations.

3. DISCUSSION

The only case where is always possible to find an identical pattern of area relationships, between the implementations, Lisbeth and Treefitter, is when the data sets have no more than one instance of ambiguity, but this condition is trivial, because real data sets are often riddled with ambiguity information, thus LisBeth and Treefitter usually uncover different patterns of relationships. But with our results, we observed that the difference between both implementations does not depend on the amount of instances of ambiguity in the data sets. Instead, the value of MS index is dependent on the position of the instance of ambiguity in the each original areagram/TAC, that is to say, the difference between both implementations is due to their theoretical dissimilarity for treatment of each type of ambiguity rather than the accumulation of noise in a data set (Table 4).

3.1 MISSING AREAS

One source of incongruence between both implementations are the missing areas. In the simulated data sets with this type of ambiguity we observed that both implementations found several different results (Appendix 2). This is because whereas the default cost values used by Treefitter, clearly favors the explanation of the missing information as being due to primitive absence and therefore the non-missing areas should be grouped (Sanmartín and Ronquist, 2002). LisBeth does not include missing areas in the resulting three-item statements (3is), because they are just treated as missing (Zaragüeta-Bagils and Bourdon, 2007). That is, each hypothesis of relationship is decomposed into 3is, but the missing terminals are not included in these resulting 3is. In one of the evaluated data set with the maximum missing information for the F* area (Fig. 9), is easy to see this behavior, whereas

Treefitter fits the area with the most missing information to the basal position because grouped the non-missing areas into a monophyletic group, for LisBeth the area with more missing information forms a internal polytomy (Fig. 9).

LisBeth differs from Treefitter as it builds a matrix of three-taxon statements whereas Treffitter uses a direct optimization procedure using parsimony. Some authors claim that there is redundancy among the three-taxon statements for a particular character (Kluge, 1993; Farris *et al*., 1995). namely the transformed matrix creates evidence that is not in the original data, hence an area with missing information is only corroborated by a redundant three-taxon statements. Zaragüeta-Bagils and Bourdon (2007) counter the claim by stating that the optimisation used by Treefitter for missing areas, is compatible with any state, and therefore they lack any empirical content because they do not forbid anything (for further discussion see, Farris, 2000; Williams, 2002; Nelson *et al*., 2004; Williams and Ebach, 2006, among many others).

There are however, cases where both implementations found the same result, even with many instances of missing areas, this depends on costs assignment and the amount of missing information for each area (not of the total of missing information on the data set, see fig. 10). Multiple linear regression analyzes showed that missing areas is a property observed that influence the maximum similarity (at a lower rate compared to the instances of paralogy and MASTs), but as we showed in table 2 there is no a perfect relationship between these two properties. This is because the similarity between the patterns obtained by both implementations is not always directly related to the amount of missing areas. Besides that the similarity, is influenced by other forms of ambiguity.

3.2 MASTs

Another source of incongruence between both implementations are the number of instances of MASTs, widely discussed in literature (see Sanmartín and Ronquist, 2002; Parenti and Ebach, 2009). In the simulated data sets with this type of ambiguity (MASTs only and MASTs+missing areas), we observed that both methods found very different results (Appendix 2). In fact, it is easier to find differences that getting a similarity between both implementations when dealing with this type of ambiguity. This is because Treefitter might assumes that the widespread taxa is the result of recent dispersal and these events are not counted when calculating the cost of the optimal GAC(s) (Sanmartín, 2007; Sanmartín *et al*., 2007), whereas LisBeth discovers the pattern of relationships derived from MASTs, through observed information by the transparent method. That is, capture what relationships are represented in each initial areagram by treated individually each MAST (Ebach *et al*., 2005). In a evaluated data set with MASTs only, we could see this behavior (Fig. 11), whereas in Treefitter, the widespread taxa are pushed further down in the GAC because only recent dispersal is allowed, (Sanmartín, 2007), LisBeth found another internal position.

In the same way, there are cases where both implementations have found the same result with instances of MASTs (but they are scarce). This is because some data sets do not have a taxon with a larger number of MASTs than the others, but rather the instances of MASTs are presented throughout the data set (Fig. 12). Multiple linear regression analyzes showed that instances of MASTs is one of the properties observed that influence the MS between both implementations, but as we showed in table 2 there is not a perfect relationship between these two properties, that is because, the similarity between the patterns obtained by both implementations, is not always directly related to the amount of instances of MASTs, but rather is dependent on the position of the instance of MASTs in the each original areagram/TAC.

3.3 PARALOGY

The final source of incongruence between both implementations is the number of instances of Paralogy. In the simulated data sets with this type of ambiguity (paralogy only and paralogy+missing), we found the highest MS values (Appendix 2). In Fact, it is easier to find similarities that getting a difference between both implementations when dealing with this type of ambiguity. This is because both implementations have a analogous form to resolve/remove this type of ambiguity which is dependent of the patristic distances between instances of paralogy. that is, the default cost values used by Treefitter, may favor both duplication + extinction or vicariance + dispersal events to resolve the instances of paralogy, this depends on the position of the ambiguity (Fig 6). In the analysis of Patristic Distances between instances of paralogy, when the pattern of area relationships was the same for both implementations, Treefitter found the events duplication + sorting to resolve the instances of paralogy, and analogously eliminated the paralogy as LisBeth. But when the patristic distance of the central paralogous area (A*) was ≥ 3 with respect to its analogous areas, Treefitter explains the history through some events of vicariance + dispersal and found different patterns of relationships than LisBeth (Fig 6b,c). This behavior is also reflected in the simulated data sets, some examples are showed in figs. 13, 14, where Treefitter found both vicariance + dispersal or duplication + extinction events, to resolved the pattern of areas relationships.

CONCLUSIONS

Each of the ambiguity types produces noise in the data sets, but the similarity or difference between the patterns of relationships found by Parsimony-based tree fitting and Paralogy-free subtrees analysis depends on the position rather than the accumulation of this ambiguity. This is just an observation because, the position of any instances of ambiguity can not be manipulated and therefore in most occasions the methods will tend to found very different results.

The results of this study reflects the large methodological differences between the methods. Such as whether to discard or utilize ambiguous information, process the data with systematic methods or event-based models or treat nodes as relations or distributions. The theoretical framework of these methods are so different that similar results are best described as coincidences. This reflects the ídentity crisis' suffered by the biogeography (Riddle, 2005). As a research program, biogeography does not share consistent theoretical or methodological aims and usually give contrasting results. This is problematic as all biogeographers are interested in 'true historical relationships, which means that there is a common theme, but with different agendas'. The need for a change is clearly evident, either by choosing a unifying method and excluding a portion of biogeographers or, by forming large working groups with specific agendas.

REFERENCES

BROOKS, D.R., McLENNAN, D.A., 2001. A comparison of a discovery-based and an event-based method of historical biogeography. J. Biogeogr. 28, 757-767.

CRACRAFT, J., 1986. Origin and evolution of continental biotas: Speciation and historical congruence within the australian avifauna. Evolution 40, 977-996.

CRISCI, J.V., 2001. The voice of historical biogeography. J. Biogeogr. 28, 157-168.

DOWLING, A.P., VELLER, M.G.V., HOBERG, E.P., BROOKS, D.R., 2003. A priori and a posteriori methods in comparative evolutionary studies of host-parasite associations. Cladistics 19, 240-253.

DUCASSE, J., CAO, N., ZARAGÜETA-BAGILS, R., 2008. Lisbeth. three-item analysis software package. Laboratoire Informatique et Systématique, UPMC Univ Paris 06, UMR 7207 (CR2P) CNRS MNHN UPMC.

EBACH, M.C., 2003 Area cladistics. Biologist 50, 169-172.

EBACH, M.C., EDGECOMBE, G.D., 2001. Cladistic biogeography: component-based methods and paleontological application. In: Adrain, J.M., Edgecombe, G.D. and Lieberman, B.S. (eds.), Fossils, Phylogeny, and Form: An Analytical Approach. Kluwer/Plenum, pp. 235-289.

EBACH, M.C., HUMPHRIES, C., WILLIAMS, D., 2003. Phylogenetic biogeography deconstructed. J. Biogeogr. 30, 1285-1296.

EBACH, M.C., HUMPHRIES, C.J., 2002. Cladistic biogeography and the art of discovery. J. Biogeogr. 29, 427-444.

EBACH, M.C., HUMPHRIES, C.J., NEWMAN, R.A., WILLIAMS, D.M., WALSH, S.A., 2005. Assumption 2: opaque to intuition? J. Biogeogr. 32, 781-787.

ESCALANTE, T., RODRÍGUEZ, G., CAO, N., EBACH, M.C., MORRONE, J.J., 2007. Cladistic biogeographic analysis suggests an early Caribbean diversification in Mexico. Naturwissenschaften 94, 561-565.

FARRIS, J.S., 2000. Diagnostic efficiency of three-taxon analysis. Cladistics 16, 403-410.

FARRIS, J.S., KALLERSJO, M., ALBERT, V.A., ALLARD, M., ANDERBERG, A., BOWDITCH, B., BULT, C., CARPENTER, J.M., CROWE, T.M., LAET, J.D., FITZHUGH, K., FROST, D., GOLOBOFF, P., HUMPHRIES, C.J., JONDELIUS, U., JUDD, D., KARIS, P. O., LIPSCOMB, D., LUCKOW, M., MINDELL, D., MUONALA, J., NIXON, K., PRESCH, W., SEBERG, O., SIDDALL, M.E., STRUWE, L., TEHIER, A., WENZEL, J., WHEELER, Q., WHEELER, W., 1995. Explanation. Cladistics 11, 211-218.

FATTORINI, S., 2008. Hovenkamp's ostracized vicariance analysis: testing new methods of historical biogeography. Cladistics 24, 611-622.

GARZÓN-ORDUÑA, I., MIRANDA-ESQUIVEL, D.R., DONATO, M., 2008. Parsimony analysis of endemicity describes but does not explain: an illustrated critique. J. Biogeogr. 35, 903-913.

HUMPHRIES, C.J., PARENTI, L.R., 1999. Cladistic biogeography: interpreting patterns of plant and animal distributions. 2nd edn. Oxford Biogeography series No. 12. Oxford University Press, Oxford.

JONG, H.D., 1998. In Search of historical biogeographic patterns in the western Mediterranean terrestrial fauna. Biol. J. Linn. Soc. 65, 99-164.

KLUGE, A.G., 1993. Tree-taxon transformation in phylogenetic inference: ambiguity and distortion as regards explanatory power. Cladistics 9, 246-259.

McLENNAN, D.A., BROOKS, D.R., 2002. Complex histories of speciation and dispersal in communities: a re-analysis of some australian bird data using BPA. J. Biogeogr. 29, 1055-1066.

MORRONE, J.J., 2009. Evolutionary biogeography: An integrative approach with case studies. Columbia University Press, Nueva York.

MORRONE, J.J., CARPENTER, J.M., 1994. In search of a method for cladistic biogeography: An empirical comparison of component analysis, brooks parsimony analysis, and tree-area statements. Cladistics 10, 99-153.

MORRONE, J.J., CRISCI, J.V., 1995. Historical biogeography: Introduction to methods. Annu. Rev. Ecol. Syst. 26, 373-401.

NELSON, G., LADIGES, P.Y., 1995. TAX: MsDos computer programs for systematics. New York and Melbourne: Published by the authors.

NELSON, G., LADIGES, P.Y., 1996. Paralogy in cladistic biogeography and analysis of paralogy-free subtrees. Am. Mus. Novit. 3167, 1-58.

NELSON, G., WILLIAMS, D.M., EBACH, M.C., 2004. A question of conflict: three-item and standard parsimony compared. Systematics and Biodiversity 1, 145-149.

PARENTI, L.R., EBACH, M.C., 2009. Comparative Biogeography: Discovering and Classifying Biogeographical Patterns of a Dynamic Earth. University of California Press, Berkeley.

POSADAS, P., MORRONE, J.J., 2003. Biogeografía histórica de la familia curculionidae (coleoptera) en las subregiones subantártica y chilena central. Revista de la Sociedad Entomológica Argentina 62, 75-84.

RIDDLE, B.R., 2005. Is biogeography emerging from its identity crisis?. J. Biogeogr. 32, 185-186.

RONQUIST, F., 2003a. Parsimony analysis of coevolving species associations. In: Page R.D.M. (ed.), Tangled Trees Phylogeny, Coespeciation and Coevolution. Chicago University Press, Chicago, pp : 22–64.

RONQUIST, F., 2003b. Treefitter, version 1.3b. Software available from http://www.ebc.uu.se/systzoo/research/Treefitter/Treefitter.html.

RONQUIST, F., NYLIN, S., 1990. Process and pattern in the evolution of species associations. Syst. Zool. 39, 323-344.

ROSEN, D.E., 1978. Vicariant patterns and historical explanation in biogeography. Syst. Zool. 27, 159-188.

SANMARTÍN, I., 2007. Event-Based Biogeography: Integrating patterns, processes, and time. In: Ebach M.C. , Tangney R.S. (eds.), Biogeography in a

changing world. Systematics Association Volume Series, Taylor & Francis, London, pp. 135-156.

SANMARTÍN, I., ENGHOFF, H., RONQUIST, F., 2001. Patterns of animal dispersal, vicariance and diversification in the Holarctic. Biol. J. Linn. Soc. 73, 345-390.

SANMARTÍN, I., RONQUIST, F., 2002. New solutions to old problems: widespread taxa, redundant distributions and missing areas in event-based biogeography. Anim. Biodivers. Conserv. 25, 75-93.

SANMARTÍN, I., RONQUIST, F., 2004. Southern hemisphere biogeography inferred by event-based models: Plant versus animal patterns. Syst. Biol. 53, 216-243.

SANMARTÍN, I., WANNTORP, L., WINKWORTH, R.C., 2007. West wind drift revisited: testing for directional dispersal in the southern hemisphere using event-based tree fitting. J. Biogeogr. 34, 398-416.

SANTINI, F., WINTERBOTTOM, R., 2002. Historical biogeography of Indo-Western pacific coral reef biota: is the indonesian region a centre of origin? J. Biogeogr. 29, 189-205.

VAN VELLER, M.G.P., BROOKS, D.R., 2001. When simplicity is not parsimonious: a priori and a posteriori methods in historical biogeography. J. Biogeogr. 28, 1-11.

VILLALOBOS, F., 2006. A performance evaluation of patter-based and event-based methods of historical biogeography: Recovering the historical signal. Metodos en Ecología y Sistematica 1, 44-62.

WILLIAMS, D.M., 2002. Precision and parsimony. TAXON 51, 143-149.

WILLIAMS, D.M., EBACH, M.C., 2006. The data matrix. Geodiversitas 28, 409-420.

ZARAGÜETA-BAGILS, R., BOURDON, E., 2007. Three-item analysis: Hierarchical representation and treatment of missing and inapplicable data. J. Syst. Palaeontol. 6, 527-534.

ANNEXES

ANNEX A. TABLES

Table 1. Total instances of ambiguity in the Australian data set. Gray row = Ambiguity estimated for Cinclosoma; Paralogy Max = Maximum instances of Paralogy for each original areagram/TAC; Paralogy Obs = Instances of paralogy observed for each original areagrams/TACs; MASTs Max = Maximum instances of MASTs for each original areagram/TAC; MASTs Obs = instances of MASTs observed for each original areagrams/TACs.

Philogenetic Trees	Paralogy Max.	Paralogy Obs.	MASTs Max.	MASTs Obs.
Psophodes	60	0	66	1
Stipiturus	36	0	44	2
Cinclosoma	60	5	66	4
Poephila	60	2	66	1
Malurus	84	4	88	2
Petrophasa	24	0	33	0
Ptiloris	28	0	55	0
Tregellasia	36	0	44	0
∑	164	10	174	10
	Total instances of Paralogy = 10/164		Total instances of MASTs = 10/174	

Table 2. The properties of each real data set and MS value found between Treefitter and LisBeth. Areas = Number of Areas in each data set; Terminals = Number of terminals in each data set; Areagrams/TACs = Number of areagrams or TACs (depending on the method) used in each data set; Paralogy = Total instances of paralogy for each data set; MASTs = Total instances of MASTs for each data set; Missing Areas = Total instances of missing areas for each data set; MS = Maximun similarity found between Treefitter and LisBeth.

Data sets	Areas	Terminals	Areagrams/TACs	Paralogy	MASTs	Missing Areas	MS
Central America	9	14	2	0	0,026	0,55	1
Subantarctic and Central Chilean	6	71	6	0,223	0,128	0,27	0,5
Indo-western Pacific	18	40	7	0,107	0,123	0,5	0
Autralia	12	42	8	0,028	0,019	0,59	0,375
western Mediterranean	13	101	9	0,09	0,073	0,33	0,1
Nearctic and Neotropical	16	287	37	0,072	0,056	0,56	0
Holartic	8	798	57	0,137	0,042	0,49	0,33

Table 3. The properties that influence the similarity between both implementations. Model = The properties that influence the similarity between both implementations; adj. R^2 = adjusted R^2.

Criterion	Model	R^2	adj.R^2
exhaustive	# areas + I. paralogy I. MASTs+ I. missing areas	0.99	0.98

Table 4. Differences between the implementations of both methods.

Features	LisBeth	Treefitter
Objective	Uses the naturally hierarchical phylogenetic relationships of clades to discover the area relationships.	Finds a pattern of historical distribution among multiples group of organisms, and to identify the causal mechanisms of this distribution over time. Include four biogeographic mechanisms into the analysis as possible explanations for the observed pattern.
treatment of Paralogy	Paralogy is removed because copies are not added anything to the analysis and may give erroneous results.	This ambiguity no need a special protocols. A priori case of duplication can be treated either as duplication + sorting (all occurrences due to ancestry), or Vicariance + dispersal (some of the occurrences possibly due to dispersal).
treatment of MASTs	This ambiguity is treated by the transparent method (Ebach et al., 2005). Hence the MASTs are treated individually so each area is represented as a terminal node.	MASTs can be treated with the Recent, ancient or free options (Some studies showed that the recent option is more powerful then other options, the reason for this is that the recent option treated the MASTs as the result of recent dispersal).
treatment of Missing Areas	Missing areas are not included in the resulting 3is, because missing data are just treated as missing in three-item analysis.	This ambiguity is treated as true absence and explain them as due to primitive absence or extintion, this may be taken as evidence that the non-missing areas should be grouped.
Algorithm	3ia-Compatibility	Parsimony
Consensus	Intersection tree	This implementation have not a form of consensus, but usually the strict consensus is applied when more than one GAC is found.
Statistical significance	NA	Significance of results can be assessed by comparing them with those derived from ramdom data set under the null hypothesis that distributions of taxa are not congruent.

ANNEX B. FIGURES

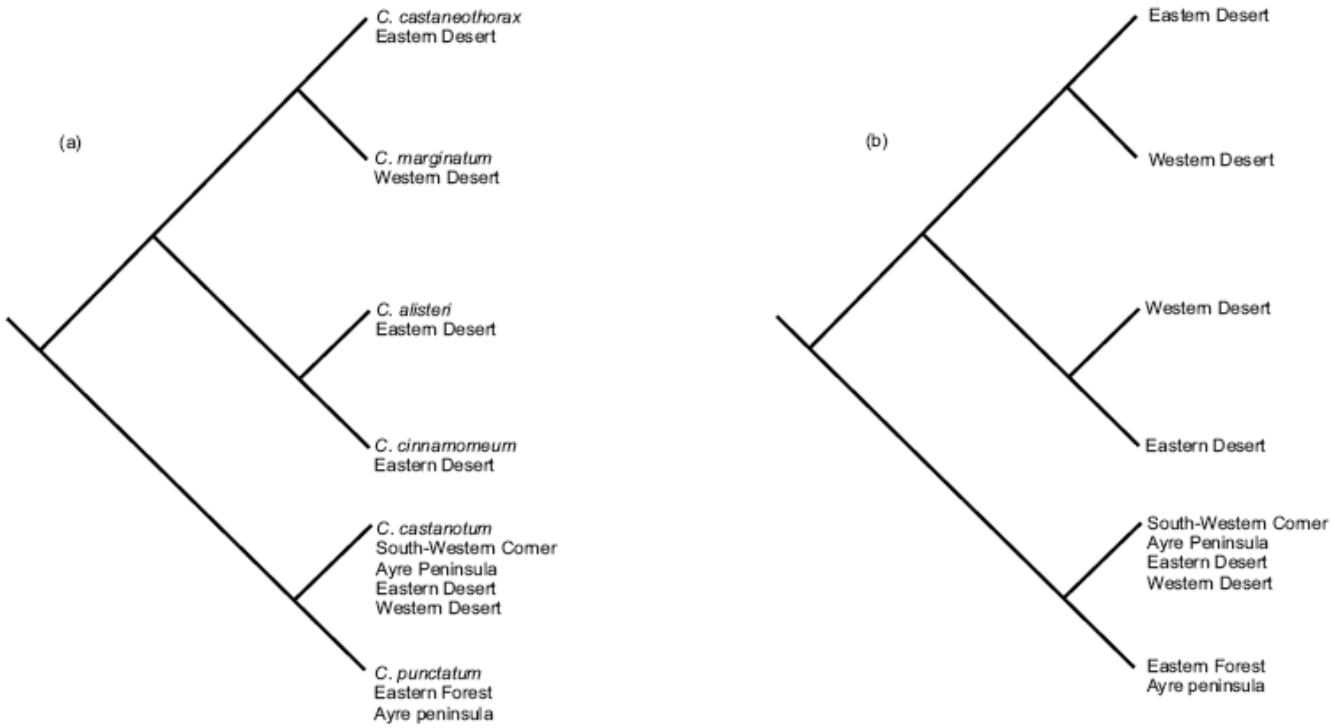


Figure 1. Original Areagram and TAC derived from *Cinclosoma* taxa. (a) TAC derived from the cladogram of *Cinclosoma* (Cracraft, 1986) by adding the endemic area in which each taxon lives. (b) Areagram of Australia and New Guinea, derived by replacing the name of the *Cinclosoma* taxa with the name of the endemic area in which it lives.

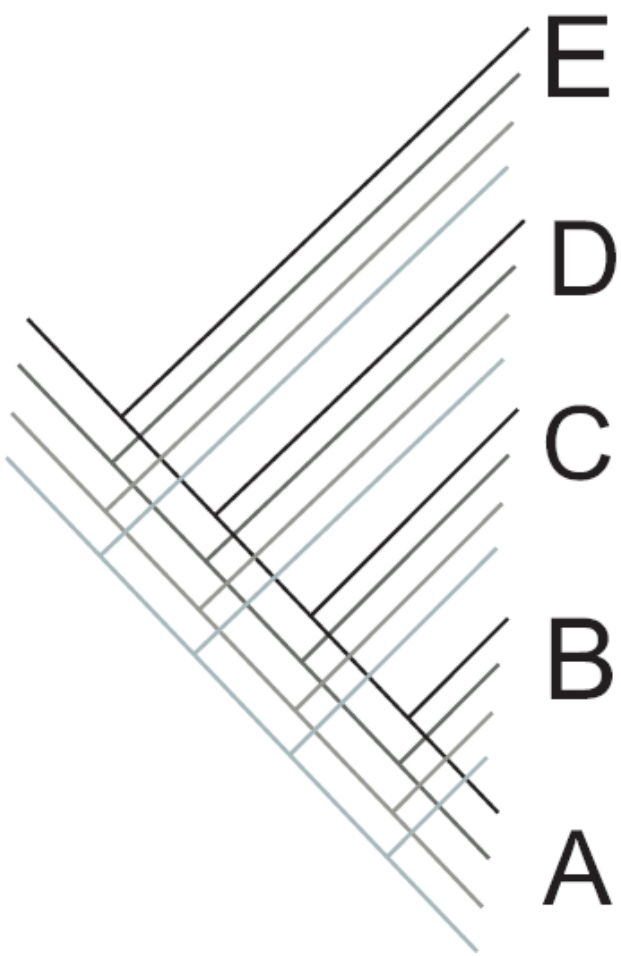


Figure 2. Data set without ambiguity instances (as negative control), hence each taxon inhabits an area, every area are only once in each original areagrams/TACs and each original areagrams/TACs has all areas of interest.

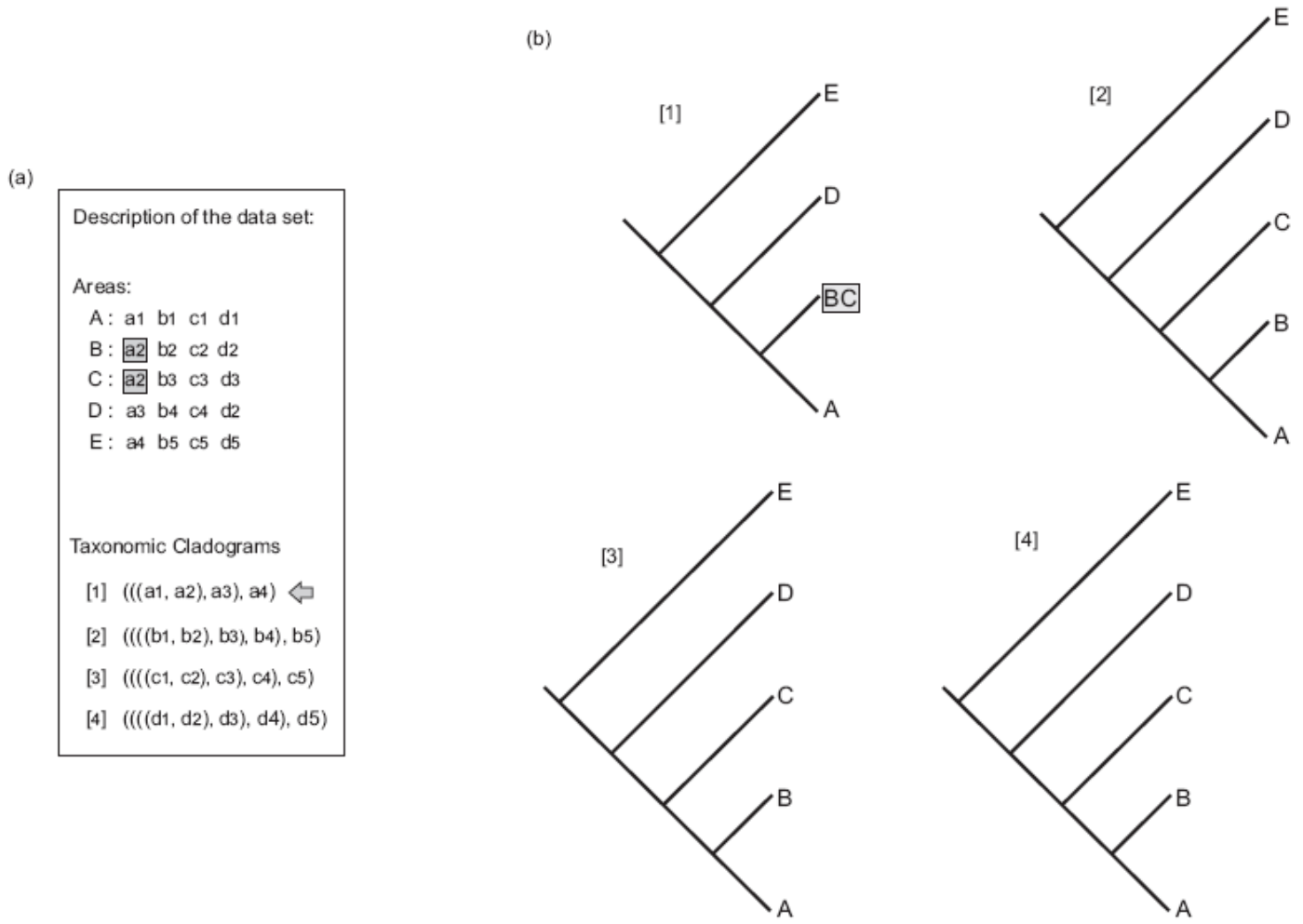


Figure 3. Data set with one instances of ambiguity, case: MASTs only, (a) Description of the data set. Gray boxes = One instance of MAST; arrow = Taxon removed from the analysis, (b) Diagram branches for each original areagram. Gray boxes = One instance of MAST.

(a)

Description of the data set:

Areas:

A :	a1	ø		c1	d1
B :	a2	a3	b2	c2	ø
C :	ø		b3	c3	d3
D :	a3	a4	b1 b4	c4	d4
E :	ø		b5	c5	d2 d5

Taxonomic Cladograms

[1] ((((a1, a2), a3), a4), a5)

[2] ((((b1, b2), b3), b4), b5)

[3] ((((c1, c2), c3), c4), c5)

[4] ((((d1, d2), d3), d4), d5)

(b)

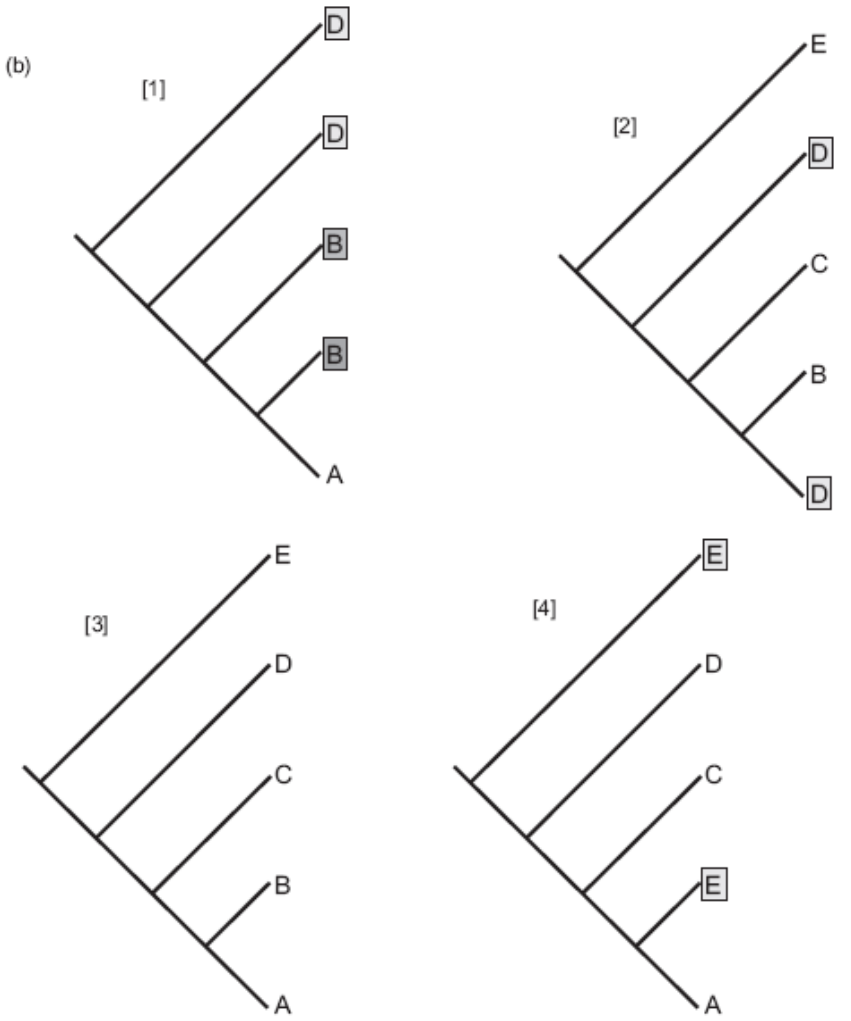


Figure 4. Data set with more than one instances of ambiguity, case: paralogy+missing areas, (a) Description of the data set. Empty symbol (ø) = Missing taxa in an area; gray boxes = One Instance of paralogy (b) Diagram branches for each original areagram. Gray boxes = instance of paralogy.

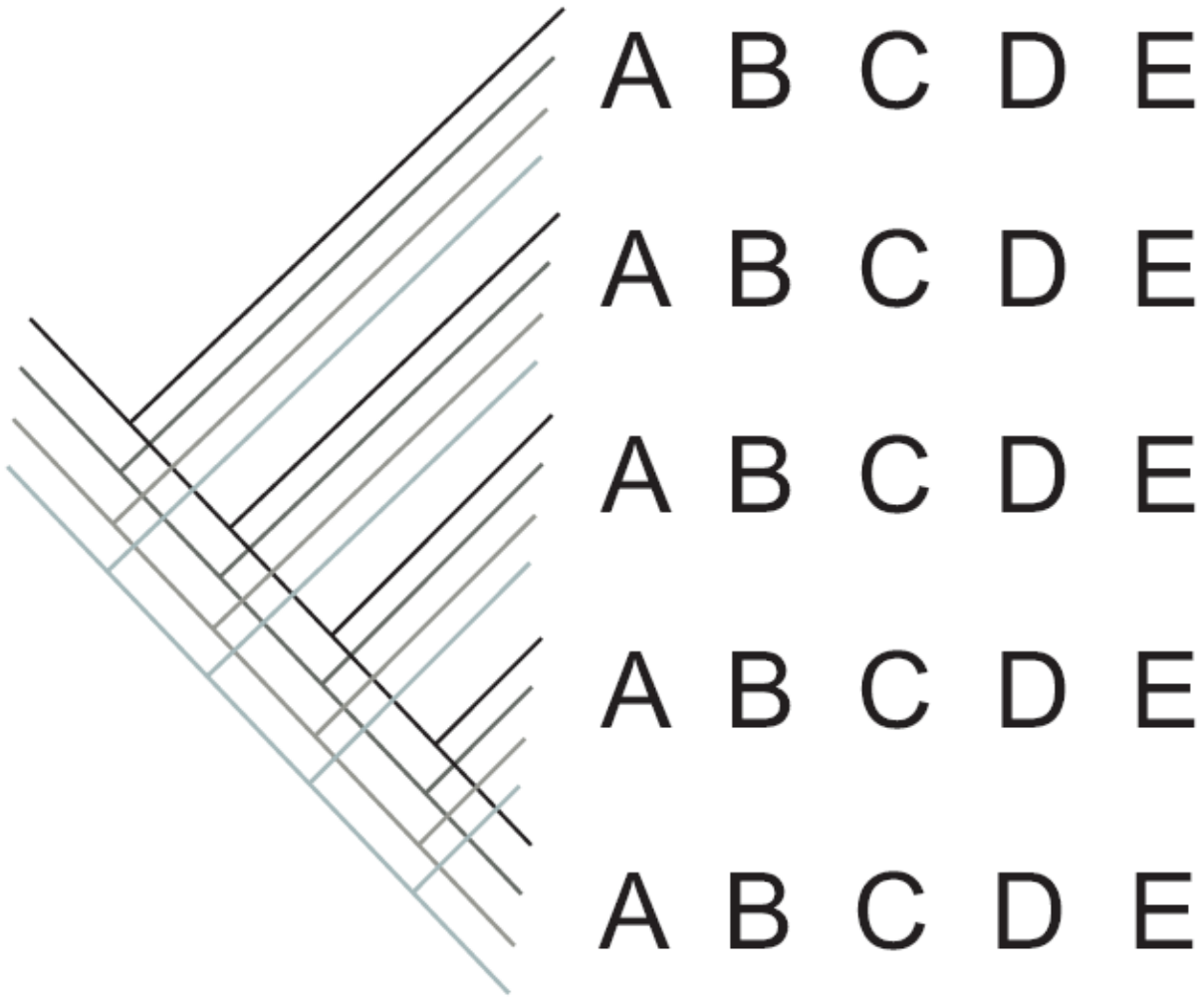


Figure 5. Data set with the maximum instances of paralogy and MASTs (as positive control), hence every taxon inhabits all areas.

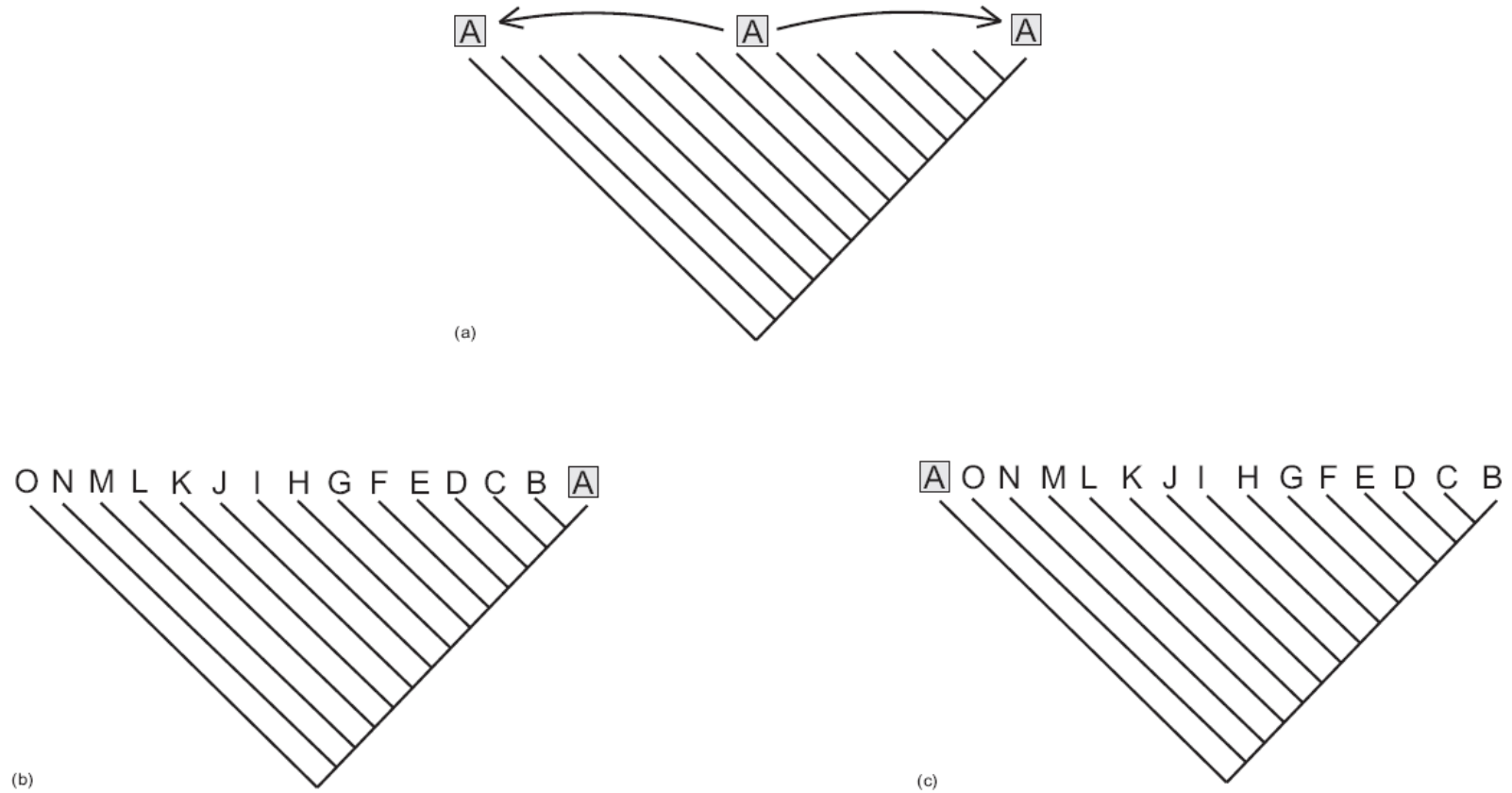


Figure 6. Patristic distances of the instances of paralogy+missing areas. (a) the central paralogous area (A*) was displaced towards ends to the patterns of relationships, one component at a time, (b) When the patristic distances was ≥ 3, LisBeth found an internal relationship to the paralogous area, (c) When the patristic distances was ≥ 3, Treefitter moves the paralogous area towards the root of pattern.

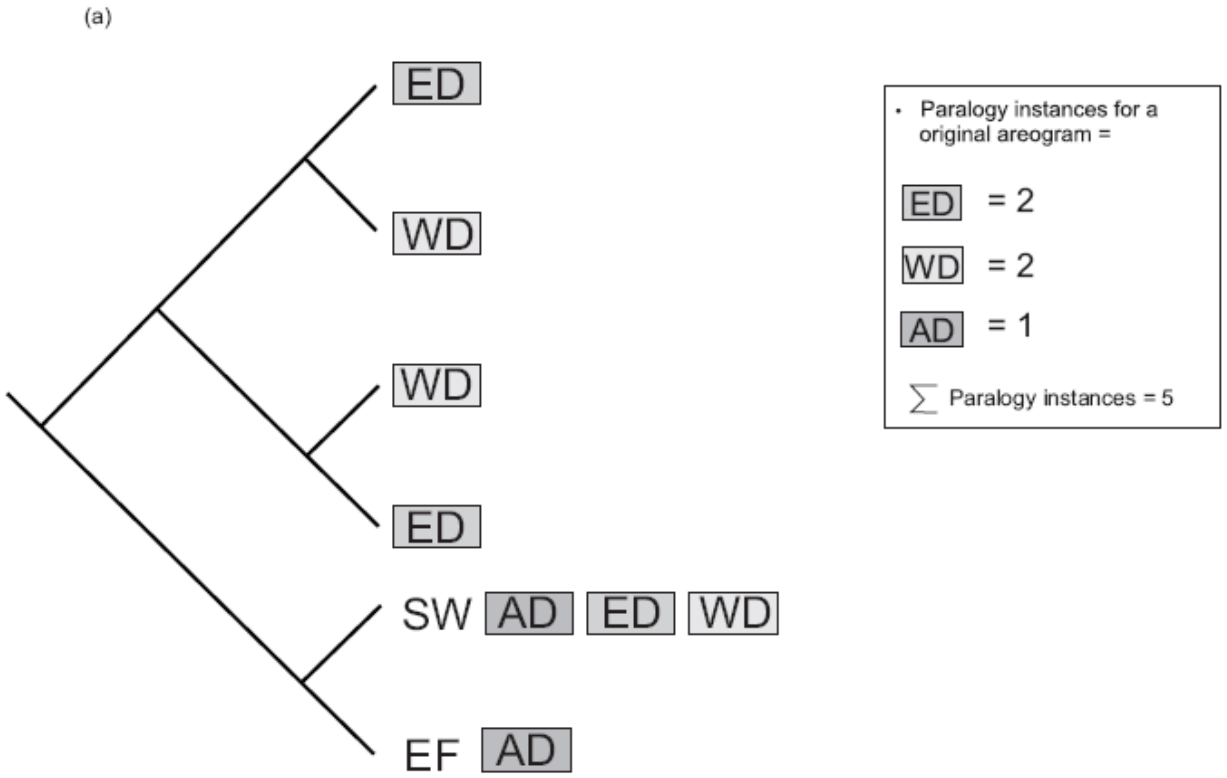


Figure 7. Calculation of instances of ambiguity observed from areagram/TAC derived from *Cinclosoma* taxa (a) Instances of paralogy observed, for the original areagram/TAC derived from *Cinclosoma* taxa. Gray boxes = Instances of paralogy, (b) Instances of MASTs observed, for the original areagram/TAC derived by *Cinclosoma* taxa. Gray boxes = Instances of MAST.

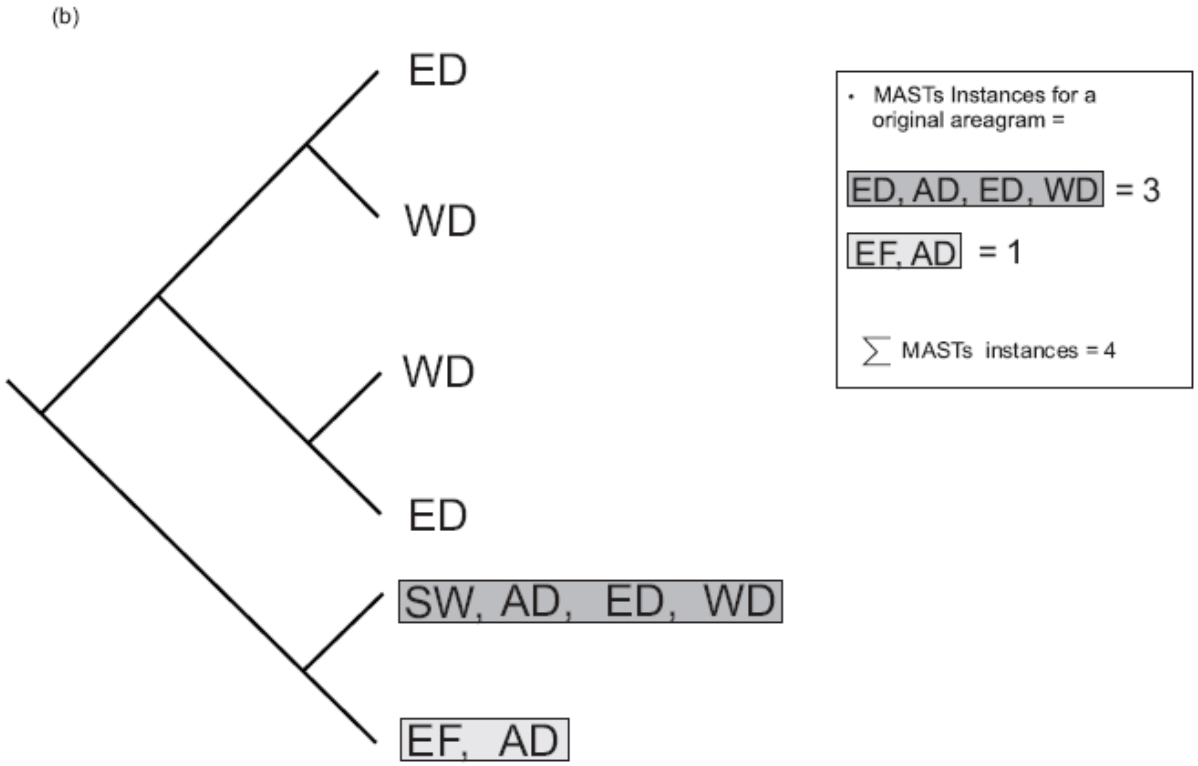
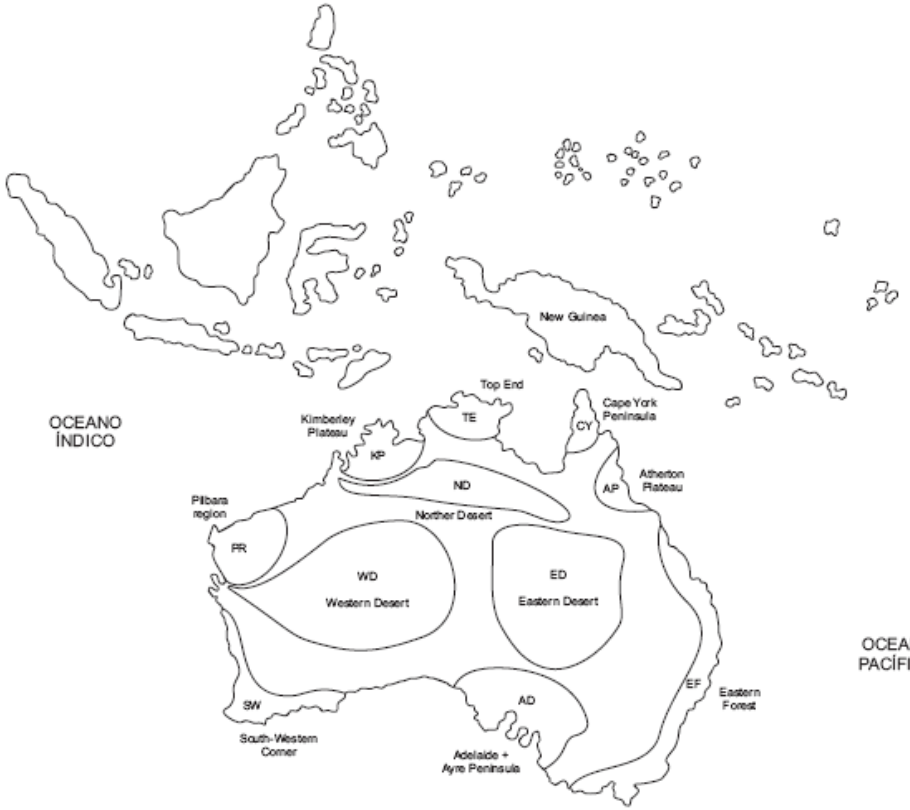


Figure 7. Calculation of instances of ambiguity observed from areagram/TAC derived from *Cinclosoma* taxa (a) Instances of paralogy observed, for the original areagram/TAC derived from *Cinclosoma* taxa. Gray boxes = Instances of paralogy, (b) Instances of MASTs observed, for the original areagram/TAC derived by *Cinclosoma* taxa. Gray boxes = Instances of MAST.



Phylogenetic Trees	Missing Areas
Psophodes	5
Stipiturus	6
Cinclosoma	8
Poephila	8
Malurus	6
Petrophasa	9
Ptiloris	8
Tregellasia	8
$\sum$	58

- Total of missing information: 58/96

- Minimum of missing information: 12 areas x 8 areagrams = 96

Figure 8. Calculation of missing information to Australian data set (a) Minimum of missing information to Australian data set, (b) missing information observed and the total of missing information to Australian data set.

(a)

Description of the data set:

Areas:

A : a1 b1 c1 d1
B : a2 ø c2 d2
C : a3 b3 c3 d3
D : a4 b4 c4 d4
E : a5 b5 c5 d5
F : ø b2 ø ø

Taxonomic Cladograms

[1] ((((a1, a2), a3), a4), a5)

[2] ((((b1, b2), b3), b4), b5)

[3] ((((c1, c2), c3), c4), c5)

[4] ((((d1, d2), d3), d4), d5)

(b)

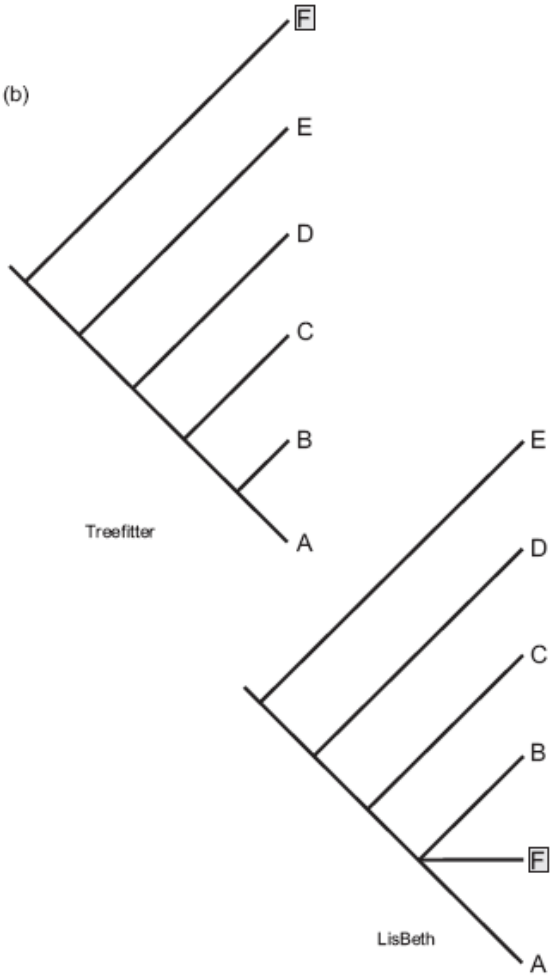


Figure 9. Example of the simulated data set, when both implementations found different results with instances of missing areas, (a) Description of the data set with maximum missing information for the F* area. Empty symbol (ø) = missing taxa in the areas, (b) Patterns of area relationships found by both implementations. Gray boxes = Conflict area between implementations.

(a)

Description of the data set:

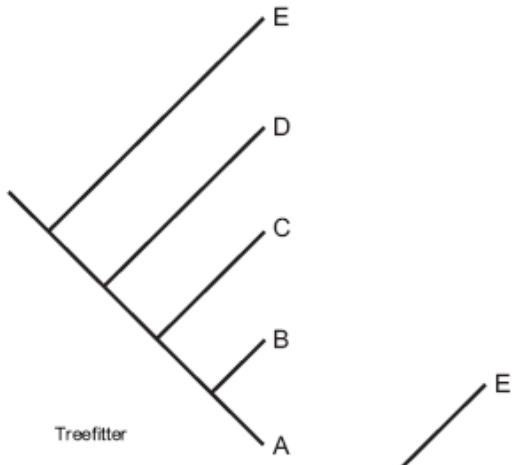
Areas:

A : ø b1 c1 d1
B : a1 b2 c2 d2
C : a2 ø ø d3
D : a3 b3 c3 ø
E : a4 b4 c4 ø

Taxonomic Cladograms

[1] (((a1, a2), a3), a4) ⇦
[2] (((b1, b2), b3), b4) ⇦
[3] (((c1, c2), c3), c4) ⇦
[4] ((d1, d2), d3) ⇦

(b)



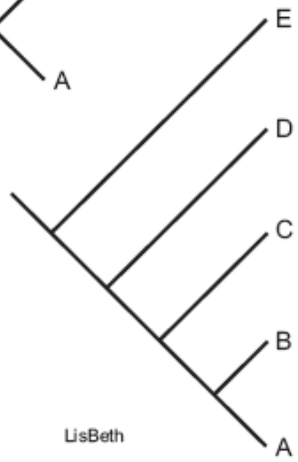


Figure 10. Example of the simulated data set, when both implementations found the same patterns of relationships with instances of missing areas, (a) Description of the data set with missing information. Empty symbol (ø) = missing taxa in the areas ; arrows = Taxa removed from the analysis, (b) Patterns of area relationships found by both implementations.

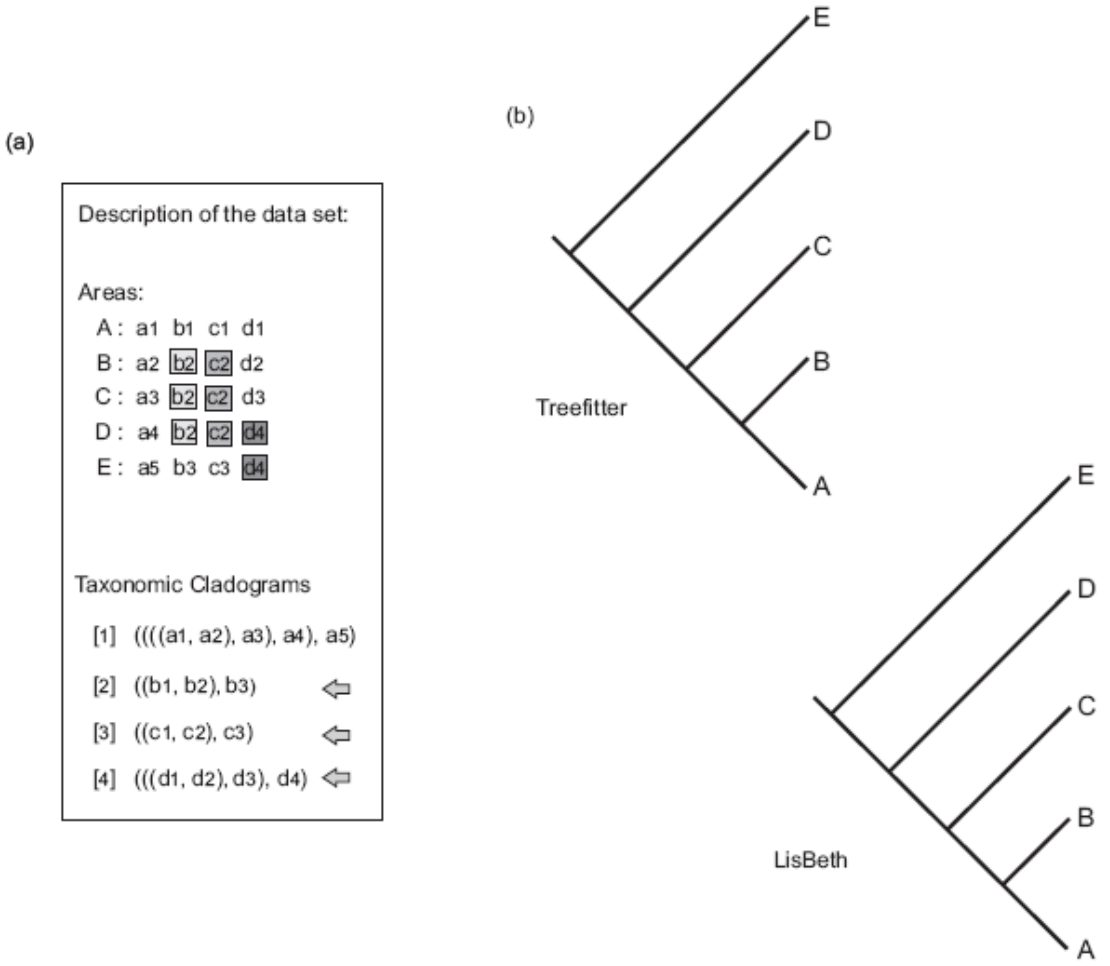


Figure 11. Example of the simulated data set, when both implementations found different results with MASTs instances, (a) Description of the data set with instances of MASTs. Gray boxes = Instances of MASTs, (b) Patterns of area relationships by both implementations, Gray boxes = Conflict area between implementations.

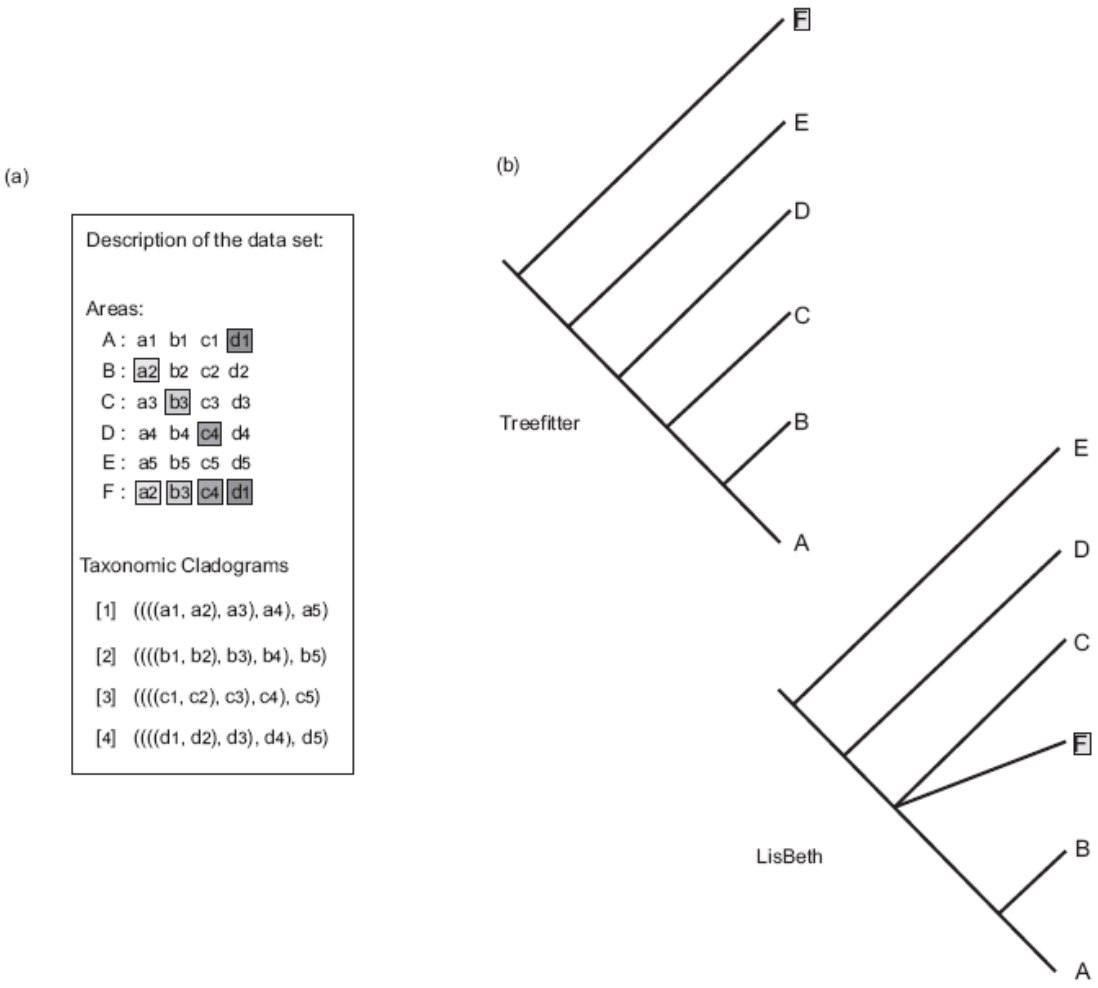


Figure 12. Example of the simulated data set, when both implementations found the same patterns of relationships with MASTs instances, (a) Description of the data set with instances of MASTs. Gray boxes = Instances of MASTs ; arrows = Taxa removed from the analysis, (b) Patterns of area relationships found by both implementations.

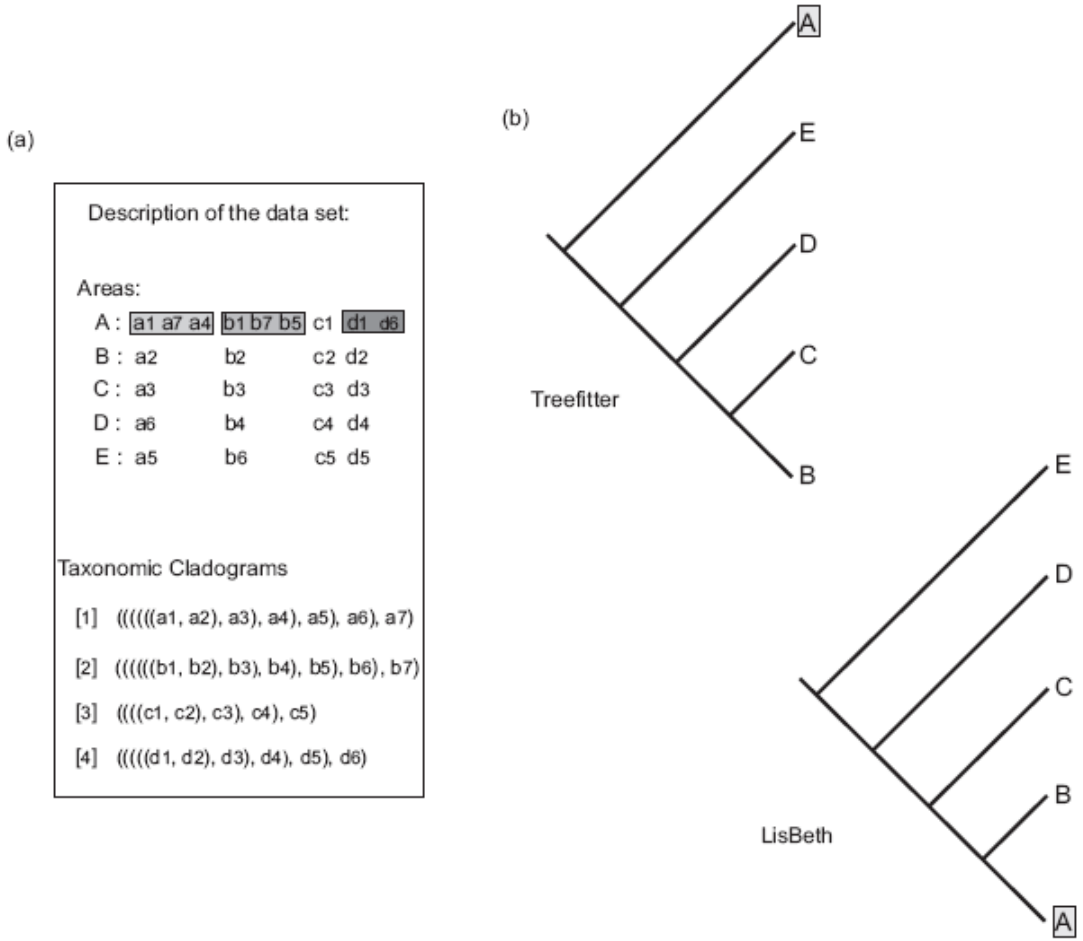


Figure 13. Example of the simulated data set, when both implementations found different results with Paralogy instances. (a) Description of the data set with instances of paralogy. Gray boxes = Instances of paralogy, (b) Patterns of area relationships found by both implementations. Gray boxes = Conflict area between implementations.

(a)

Description of the data set:

Areas:

A :	a1	b1	c1 c8	d1
B :	a2	b2	c2	d2
C :	a3 a6	b3 b6	c3	d3
D :	a4	b4	c4	d4 d6
E :	a5	b5	c5 c6 c7	d5

Taxonomic Cladograms

[1] (((((a1, a2), a3), a4), a5), a6)

[2] (((((b1, b2), b3), b4), b5), b6)

[3] (((((((c1, c2), c3), c4), c5), c6), c7), c8)

[4] (((((d1, d2), d3), d4), d5), d6)

(b)

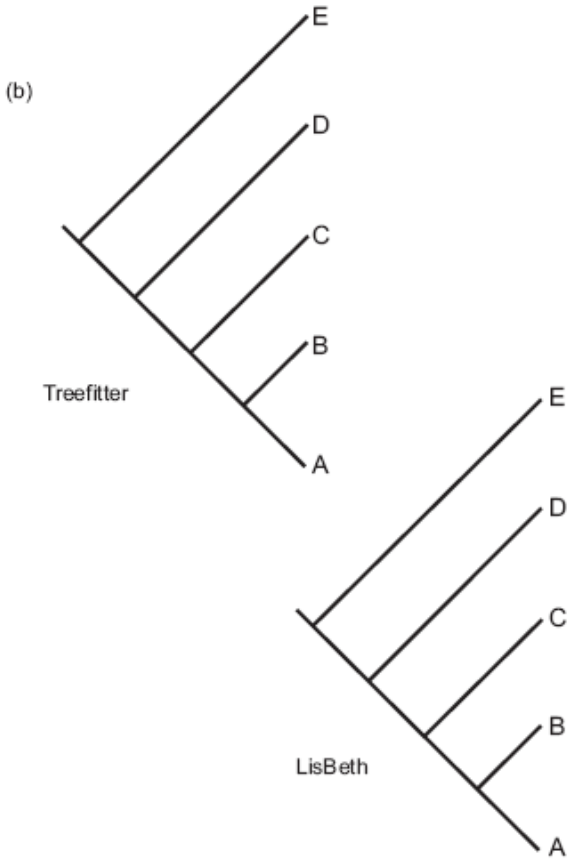


Figure 14. Example of the simulated data set, when both implementations found the same patterns of relationships with Paralogy instances. (a) Description of the data set with instances of paralogy. Gray boxes = Instances of paralogy, (b) Patterns of area relationships found by both implementations.

ANNEX C. APPENDIX

Appendix 1. Analysis with one instance of ambiguity. Cases = Different types of ambiguity analyzed; Dataset = Data sets in each case; paralogy = Instances of paralogy in each data set; MASTs = Instances of MASTs in each data set; missing areas = Instances of missing areas in each data set; MS = maximum similarity index between LisBeth and Treefitter; Observations = Differences between the data sets and the negative control.

Cases	Dataset	Instances of ambiguity				Observations
		paralogy	MASTs	missing areas	MS	
paralogy only	1	1	0	0	1	One terminal is added at random
	2	1	0	0	1	
	3	1	0	0	1	
	4	1	0	0	1	
	5	1	0	0	1	
MASTs only	1	0	1	0	1	One terminal is removed at random
	2	0	1	0	1	
	3	0	1	0	1	
	4	0	1	0	1	
	5	0	1	0	1	
missing areas only	1	0	0	1	1	One terminal is added at random
	2	0	0	1	1	
	3	0	0	1	1	
	4	0	0	1	1	
	5	0	0	1	1	

Appendix 2. Analysis with more than one instance of ambiguity. Cases = Different types of ambiguity analyzed; Dataset = data sets in each case; paralogy = Instances of paralogy in each data set; MASTs = Instances of MASTs in each data set; missing areas = Instances of missing areas in each data set; MS = maximum similarity index between LisBeth and Treefitter; Observations = differences between the data sets and the negative control.

Cases	Dataset	Instances of ambiguity				Observations
		paralogy	MASTs	missing areas	MS	
paralogy only	1	3	0	0	1	Two or more terminals are added at random
	2	4	0	0	1	
	3	5	0	0	1	
	4	3	0	0	1	
	5	7	0	0	1	
	6	6	0	0	1	
	7	7	0	0	0	
	8	8	0	0	0	
	9	7	0	0	1	
	10	5	0	0	0	
MASTs only	1	0	8	0	0.5	One or more areas are added or two or more terminals are removed at random
	2	0	3	0	1	
	3	0	5	0	0.33	
	4	0	2	0	1	
	5	0	4	0	1	
	6	0	4	0	0.33	
	7	0	4	0	0.33	
	8	0	3	0	1	
	9	0	5	0	1	
	10	0	6	0	0.33	
missing areas only	1	0	0	4	0	One or more areas are added or two or more terminals are removed at random
	2	0	0	3	1	
	3	0	0	5	1	
	4	0	0	5	1	
	5	0	0	4	1	
	6	0	0	8	0	
	7	0	0	8	0	
	8	0	0	3	1	
	9	0	0	4	0	
	10	0	0	5	1	

Cases	Dataset	Instances of ambiguity				Observations
		paralogy	MASTs	missing areas	MS	
paralogy+missing areas	1	2	0	1	1	Distributions of some taxa are moved or one or more areas are added
	2	2	0	2	1	
	3	4	0	2	1	
	4	2	0	4	1	
	5	5	0	2	1	
	6	8	0	5	1	
	7	2	0	8	0.5	
	8	3	0	7	0.25	
	9	4	0	3	0.66	
	10		0	4	1	
MASTs+missing areas	1	0	1	3	0.66	One or more areas are added or two or more terminals are removed at random
	2	0	3	1	0.5	
	3	0	5	3	0.25	
	4	0	7	1	0.5	
	5	0	3	5	0.25	
	6	0	7	5	0.5	
	7	0	3	1	1	
	8	0	2	2	1	
	9	0	3	1	1	
	10	0	5	2	1	
paralogy+MASTs	1	2	2	0	1	Distributions of some taxa are moved
	2	6	6	0	1	
	3	5	5	0	1	
	4	3	3	0	1	
	5	2	2	0	1	
	6	4	4	0	1	
	7	7	7	0	1	
	8	2	2	0	1	
	9	10	10	0	1	
	10	10	10	0	1	
	11	21	21	0	0.33	
	12	20	20	0	0.33	
	13	20	20	0	0.66	
	14	28	28	0	1	
	15	37	37	0	0.66	
	16	44	44	0	0	
	17	53	53	0	0	
	18	62	62	0	0	
	19	76	76	0	0	
	20	79	79	0	0	