Experimental systematics: sensitivity of cladistic methods to polarization and character ordering schemes

Valentin Rineau¹, Anaïs Grand², René Zaragüeta³, Michel Laurin¹, ⁴

¹ UMR 7207 CR2P, Sorbonne Universités, CNRS/MNHN/UPMC, Muséum national d’Histoire naturelle, Bâtiment de Géologie, Case postale 48, 43 rue Buffon, F-75231 Paris cedex 05, France
² Service du patrimoine naturel, Muséum national d’Histoire naturelle, Bâtiment reptiles et amphibiens, 25 rue Cuvier, 75005 Paris, France
³ UMR 7205 ISYEB, CNRS/MNHN/UPMC/EPHE, Muséum national d’Histoire naturelle, Bâtiment de Géologie, Case postale 48, 43 rue Buffon, F-75231 Paris cedex 05, France
⁴ E-mail: laurin@mnhn.fr

Key words: branch length, evolutionary model, hierarchy, ordered characters, outgroup polarization, parsimony, reversals, three-taxon analysis, three-taxon statements, unordered characters

Abstract

Phenotypic characters are essential to study the evolution of extant and extinct life forms and to reconstruct the tree of life. Inside the cladistics theory, parsimony is used by a large majority of systematists working on phenotypic characters, whereas 3ta is much less widespread but has triggered important debates. Many important differences in the interpretation of the cladistic theory exist between these methods, e.g. meaning and treatment of reversals, character representation as ‘data-matrices’ in parsimony (ordered and unordered), and as rooted trees (hierarchies) in 3ta. Although 3ta has received severe criticism, mostly focused in the use of software intended to be used in parsimony, only a few empirical studies have compared these methods so far. We present the results of simulations of the evolution of phenotypic traits under a Brownian motion model to characterize differences in sensitivity between parsimony and 3ta to (1) outgroup branch length, which affects the reliability of ancestral character state estimates, (2) character state ordering scheme, and (3) ingroup branch lengths that reflect the geological age of studied taxa. Our results show that the ‘nihilistic’ attitude of leaving multistate characters unordered when criteria to order are available (e.g., similarity, ontogeny, etc…) can decrease resolving power of the method (by 13.4% to 29.3%) and increase the occurrence of artefactual clades (by 5% to 15.6%). Increasing outgroup branch length significantly decreases resolving power and increases artefactual resolution, at least for paleontological trees. All simulations show that ordered parsimony is always superior to 3ta in tested parameter space. These results depend on the assumption in parsimony that reversals (as implied by the Brownian motion, as in most other models) can be evidence for the support of a clade a posteriori from an analysis or a priori on simulations with a known pattern. We discuss implications of these points of view compared to the assumption inherent in 3ta (i.e., that reversals should not support a clade as other synapomorphies do) on evolutionary models.

Contents

Introduction ...................................................................................... 129
Material and methods ................................................................. 135
Simulated character sets .............................................................. 135
Tree searches .................................................................................. 137
Tree comparisons .......................................................................... 137
Testing the results .......................................................................... 137
Results ............................................................................................. 138
Discussion ....................................................................................... 139
Brownian motion in cladistics ..................................................... 139
Reversals in phylogenetics ............................................................ 139
Uncertain polarization .................................................................. 143
Character states and ordering schemes ........................................ 143
Tree shape and branch length ....................................................... 143
Implications on simulation-based studies and evolutionary models ............................................................................. 144
Conclusion ...................................................................................... 145
Acknowledgements ......................................................................... 146
References ....................................................................................... 146

Introduction

The reconstruction of the Tree of Life is one of the most important quests in natural sciences, and it is essential for various biological and geological fields. Phylogenies are widely used in comparative and evolutionary biology (Hennig, 1966; Felsenstein, 1985) and even in conservation biology (Faith, 1992). Even though most contemporary phylogenetic studies of extant taxa are based on DNA sequence characters, phenotypic traits continue to play an important role because molecular sequences are unavailable for most long-extinct organisms, given that
the oldest ‘ancient’ DNA extracted so far is less than 1 Ma old (Callaway, 2011). According to various esti-
mates of extant and extinct biodiversity, and depending on how species are conceptualized, approximately 99% to 99.9% of life forms that have existed on Earth are now extinct (Miller and Foote, 1996; Newman and Sibani, 1999; Alroy, 2001). Thus, it is important to con-
tinue using and improving methods of phylogenetic in-
ference from phenotypic data. The dominant phyloge-
netic theory for phenotypic characters is cladistics. This theory is based on the assumptions that the ep-
omic access we have to discover monophyletic taxa is
through phylogenetic arguments called characters.
Characters are based on hypotheses of homology, i.e.
hypotheses that relate the observation of different fea-
tures as being the same. This identity is explained by
their evolutionary origin as parts evolved from the
same part in the exclusive (last) common ancestor.
The method of argumentation is different in parsimony
analysis and three-taxa analysis (hereafter, 3ta). In
parsimony analysis, characters-states transform into
one another, which is consistent with Hennig’s original
description of cladistic theory. In 3ta, characters-states
differentiate from one another, which is consistent with
the relative difficulty of taxa represented as hierarchies: if
homologs are parts of taxa and taxa differentiate, then
character-states should also differentiate. Theoretically,
the criterion of choice among contradictory hypotheses
is maximizing congruence. Congruence has been inter-
preted as minimizing a particular tree-distance (in par-
simony method) or as maximizing compatibility
among characters (as in the compatibility method; not
tested in this paper). In 3ta, the search for optimal trees
maximizing parsimony or compatibility of 3ts gives ex-
ceptly the same results (Wilkinson, 1994): this shows
that maximizing the treatment of characters is con-
tained in the method of treatment of homology and
homoplasny of 3ta and parsimony, and that the representa-
tion of characters as matrices, graphs or hierarchies is
only a consequence of this treatment.

According to various authors (Chappill, 1989; Ste-
vens, 1991; Thiele, 1993; Wiens, 2001), most morpho-
 logical characters are fundamentally quantitative. Con-
tinuous characters are usually discretized because most
cladistic software cannot handle them directly. Several
methods have been proposed for discretizing charac-
ters, such as segment coding (Simon, 1983; Farris,
1990) and gap coding (Mickeyevich and Johnson, 1976;
Almeida and Bisby, 1984; Archie, 1985) to limit the arbitrariness of the state delimitation, even though
these do not usually capture all the phylogenetic infor-
mation (Laurin and Germain, 2011) unless ordering
schemes are used to order the series of states or set
transition costs, and a distinct state is recognized for
each taxon (Wilkinson, 1992; Wiens, 2001). Thus, the
coding, as a formalization of character hypothesis of
the analysis (ordination, discretization, etc…) is of pri-
mary importance for phylogenetic inference. Charac-
ters can be separated into at least two classes (Fig. 1).
Fig. 1. Three examples of character representation: ordered in
parsimony (A; additive) andunordered in parsimony (B; non-
additive) form the class of parsimony characters, associated with
their stepmatrix (Sankoff and Rousseau, 1975) necessary in a
matrix representation, and the other classes of character trees,
hierarchies treated in 3ta (C). A) symbolizes ordered character
states, where transformation costs are given by the stepmatrix to
the right. B represents unordered states as an unrooted tree and
where all transformations between states have the same cost and
C represents a character tree, were nodes are homology hypoth-
eses under the hierarchy. The differences in their representations
(graph, unrooted tree, hierarchy) traduce both their content and
way in which corresponding methods treat them; these they
deal with incongruence they can generate; for example, in parsi-
mony graphs of Figs 1A and B, a step can correspond to trans-
formation from 1 to 2 or from 2 to 1, which makes no sense for
cladistic representation 1C, symbolizing a 3ta treatment.

Fig. 1. Three examples of character representation: ordered in
parsimony (A; additive) and unordered in parsimony (B; non-
additive) form the class of parsimony characters, associated with
their stepmatrix (Sankoff and Rousseau, 1975) necessary in a
matrix representation, and the other classes of character trees,
hierarchies treated in 3ta (C). A) symbolizes ordered character
states, where transformation costs are given by the stepmatrix to
the right. B represents unordered states as an unrooted tree and
where all transformations between states have the same cost and
C represents a character tree, were nodes are homology hypoth-
eses under the hierarchy. The differences in their representations
(graph, unrooted tree, hierarchy) traduce both their content and
way in which corresponding methods treat them; these they
deal with incongruence they can generate; for example, in parsi-
mony graphs of Figs 1A and B, a step can correspond to trans-
formation from 1 to 2 or from 2 to 1, which makes no sense for
cladistic representation 1C, symbolizing a 3ta treatment.

Letters and numbers (0-5, Z-W) represent named nodes as ho-
mologs.
ence (Wilkinson, 1995) of the different columns of the matrix does not apply to the different 3ts derived from the same character. On the contrary, FW reflects the logical mutual dependency (Williams and Siebert, 2000), as shown on Fig. 2B (but see Wilkinson et al., 2004).

With perfectly congruent data, parsimony and 3ta give always the same tree. Operationally, the only difference between them is how they deal with incongruence in the data (e.g. in Nelson et al., 2003). The history of wings in stick insects (phasmids) highlighted by Whiting et al. (2003) can illustrate the differences and similarities between parsimony and 3ta (Fig. 3). The cladogram of phasmid shows evidence for many losses and independent reacquisitions of a complex structure (Fig. 3A), the pterygote wing (Fig. 3B). In Fig. 3C, this hierarchy (character tree drawn a posteriori from the analysis) is simplified so as to illustrate the different points of view. In parsimony analysis, the character ‘wing’ is traditionally coded as binary with the states absent and present. In 3ta, the character is coded as an inclusive hierarchy with the state present nested within the states absent (Fig. 3A). This ‘absent’ and ‘present’ are both considered as complementary homologous (de Pinna, 1991) by parsimony proponents whereas only the acquisition is considered to be a homologous by the 3ta proponents (Cao et al., 2007). The class ‘absence’ includes ‘everything else’ and is not considered as a homologous (the plesiomorphic state in 3ta). 3ta proponents consider ‘homologs’ the nodes of a character state tree (Fig. 1C, 3C). For instance, in Fig. 1C, W is a homolog (i.e., states 3 and 4 are considered homologous), X is another homolog (i.e., 2, 3 and 4 are homologous), etc. The character tree shows ‘polarized’

Fig. 2. Fractional weighting (FW) and dependence among 3ts derived from the same character. A. A character relating free terminal taxa is decomposed into its elementary relationships, expressed as a list of 3ts. The decomposition results into a list of six 3ts. However, only two of the three 3ts (e.g. (A (C D)) and (A (C E))) that relate three terminals (C, D, E) to other, more distant ones (e.g. A) are necessary to build the initial hierarchical character, as shown in the set diagram. FW assigns an information content/weight of 2/3 to each of the two groups. B. A multistate character is analysed into components and then into 3ts. 3ta takes into account the fact that each component is not independent. Thus, the 3ts derived from different components of the same character may be redundant. Consider the 3ts (A (D E)): it appears in the three components. However, it represents a single node in the original character. The 3ts derived from the third component have a FW=1, i.e. they are non-redundant. The same 3ts in the first and second components are thus redundant and can be removed. The FW of the remaining 3ts of the component is then recalculated from FW=2/3 to FW=1 in the second component, and FW<1 in the first, etc... Note that the matrix treatment of the original character with their status of dependent or independent, contrary to characters in parsimony analysis.

Fig. 3. An example of character history according to cladistics. A is a modified phylogeny from Whiting et al. (2003) with transformations (obtained using parsimony under acctrain optimization) including reversals (denoted by a ‘+’ sign) and acquisitions or reacquisitions (‘*’) from a binary character absence/presence the 3ta character tree is represented as a hierarchy). The degree of gains and losses is denoted by numbers (1 for primary, 2 for secondary, etc.) in the circles placed at the corresponding nodes. The degree of the resulting state (presence or absence, appearing in parentheses below and to the left of the symbols for gains or losses) is denoted by the asterisks; primary gains or losses have none; secondary gains or losses have one; tertiary ones have two, etc. B shows the wings of Phasmatodea Leptynia hispanica (Bolivar, 1878), Phyllium bicoloratum (Gray, 1832) with wings and Pseudophasma acanthonotum (Redtenbacher, 1906) with wings and Pseudophasma acanthonotum (Redtenbacher, 1906) with convergently re-acquired wings (copyrights, respectively; Fritz Geller-Grimm and Felix Grimm - CC BY SA 3.0. Drigg, GFDL, and Driggs – PD-self). C represents with a character tree the history reconstructed a posteriori from the analysis (with the mapping of the character on the optimal tree) of the character ‘wing’ of phasms. Each node represents a parsimony apomorphy, contrary to the 3ta for which the only apomorphy is the node 1 (as only secondary homology because the only primary homology postulated was the homology between all wings).
The true ancestral state is 0, and 1 is a true synapomorphy of D, E and F, but the outgroup criterion leads to an erroneous ancestral condition. Results of the empirical studies of Grand et al. supported by parsimony and 3ta had the least resolving power and that it had greater artefactual resolution than ordered parsimony. They also clearly demonstrated that, under the conditions investigated in the simulations, unordered parsimony is more difficult to predict, but a greater accuracy of ordered parsimony could be expected for model-based simulations that allow genuine reversals. In this study, we test various hypotheses in order to refine characterization of the behaviour of parsimony (ordered and unordered) and 3ta on discretized continuous characters. Thus, we test the following hypotheses:

1. Given that our simulated data are intrinsically continuous, we expect that ordered methods (parsimony and 3ta) will outperform unordered parsimony. The relative performance between 3ta and ordered parsimony is more difficult to predict, but a greater accuracy of ordered parsimony could be expected for model-based simulations that allow genuine reversals.

2. We expect that errors in character polarity affect differently the methods, with the performance of 3ta deteriorating faster with increasing polarity errors than ordered parsimony because 3ta uses rooted character state trees, whereas ordered parsimony uses unrooted character state trees (Fig. 1). The transformations are multidirectional in parsimony, contrary to 3ta, which should result in more synapomorphies being recognized in parsimony. 3ta might be more affected by polarization errors because 3ta disregards reversals as synapomorphies, and if a character is wrongly polarized, a genuine synapomorphy would be disregarded (Fig. 4). In parsimony, if the polarity is wrong, all changes would be used, even though their interpretation might be erroneous (Fig. 4). The performance of unordered parsimony with polarization errors is more difficult to predict. These variable levels of polarization errors are obtained by changing outgroup branch lengths.

3. Tree balance and branch lengths of the real ingroup phylogeny have an effect on the behaviour of the methods: performance of phylogenetic methods can be influenced by terminal and internal branch length and tree symmetry. Based on the ratios of terminal to internal branch lengths (which can be considered a ratio of homoplasy to synapomorphy), then for ultrametric trees, performance should be best for the symmetrical tree, which has the lowest terminal/internal branch length ratio, and worst for a fully pectinate tree, which has the highest ratio. This should prevail in all three methods (throughout this paper, the ‘three methods’ will always refer to ordered parsimony, unordered parsimony, and 3ta). If we consider 3ta as the minimal phylogenetic information (or if we use a metric, such as ITRI, which relies on 3ta), we also expect that a long branch leading to a clade of several taxa is more important than a long branch length leading to a clade of two taxa, because fewer 3ts are involved in the latter. Thus, the structure of a phylogenetic tree may be expected to have an impact.

Material and methods

Simulated character sets

Reference topologies. Data was simulated on three topologies of 20 ingroup OTUs and one outgroup OTU and a total of 18 branch length settings (Fig. 5). One of the reference topologies is fully pectinate (i.e., fully asymmetrical), a second is fully symmetrical for the ingroup (balanced) and a third is an equiprobable (randomly generated) tree, intermediate in asymmetry between the two others (Fig. 5A-F). In this paper, branch lengths represent evolutionary time, but given that we simulated characters using Brownian motion (see below), branch lengths also reflect expected character variance (Felsenstein, 1985).

Reference trees. Three outgroup branch lengths were used on each of the reference topology, leading to nine ultrametric reference trees (Fig. 5A-C). Nine other branch lengths differing in both ingroup and outgroup were specified on the equiprobable tree with taxa of various geological ages, leading to 9 additional non-ultrametric reference trees (Fig. 5D-G). For six out of these nine trees, all terminal and internal branch lengths of the ingroup were set to one, and the outgroup branch was set at zero, one, three, five, six, tenths, or the full tree depth (Fig. 5D). Three of these nine branches were obtained by modifying the ingroup branch lengths on the equiprobable tree with the outgroup branch set at zero length (actual ancestor).

Simulated matrices. From each of the nine reference trees illustrated in Fig. 5A-C, 100 matrices of 100 characters × 10 taxa were simulated (i.e., a total of 9000 matrices). Similarly, 100 matrices of 100 characters × 10 taxa were simulated for each of the nine palaeontological trees illustrated in Fig. 5 D-G (i.e., a total of 9000 palaeontological matrices). We thus simulated a total of 1800 matrices, which were produced using Mesquite and the scripts in Supplementary Online Materials 1 (S1) on data that were discretized using Excel spreadsheets (S2 for parsimony; S3 for 3ta) and that were compiled into S4 (which are the matrices in the parsimony format).

Character coding. The characters were simulated with continuous Brownian motion in Mesquite (Madison and Maddison, 2014) to represent data inherently ordered as morphoclines (such as size or shape characters). Simulations were made using this evolutionary model because it is one of the simplest and most widely used in evolutionary biology to study the evolution of continuous phenotypic characters. For example, phylogenetic independent contrasts (Felsenstein, 1985) and squared-change parsimony (Maddison, 1991) assume this model. Characters simulated through Brownian motion are continuous; they were then discretized into 10 equal intervals representing character states, in order to simulate morphoclines following the simple procedure described in Laurin and Germain (2011). Because Brownian motion has no tendency, the resulting distribution is Gaussian; thus, gap coding cannot be used, and the limits between states are arbitrary. The primitive condition is determined by the outgroup criterion. The variable outgroup branch lengths allow us to assess the influence of polarization errors, whereas the variable ingroup branch lengths allow assessment of the impact of geological age of ingroup taxa on tree resolution (presumably by altering support of the clade subtended by the various branches), thus enabling a comparison of palaeontological and neontological data.
sets. Each of the 100 characters (in the 1800 matrices) was coded in three different ways corresponding to unordered parsimony, ordered parsimony and three-item analysis (3ta).

**Tree searches**

Characters were analysed with the three methods. Thus, 5400 phylogenetic analyses were performed (three analyses for each of the 1800 matrices). Cladistic analyses were performed with PAUP\$ 4.0b10 (Swofford, 2003) for both ordered and unordered parsimony. Three-item analyses were also performed on PAUP\$ 4.0b10 but parsimony matrices were transformed into three-item statements matrices with fractional weighting (Mickeyvich and Platnick, 1989; Nelson and Ladiges, 1992) using LisBeth 1.3 (Zaragüeta et al., 2012). All analyses were done with a heuristic search of 50 replicates (a number of searches that our preliminary analyses suggested was sufficient to recover all optimal trees for our matrices), using the TBR algorithm. A strict consensus (Sokal and Rohlf, 1981) of all optimal trees was then constructed using PAUP\$.

**Tree comparisons**

The optimal tree yielded by the data analyses (or strict consensus tree when several optimal trees were found) were compared with the reference phylogenies, to assess accuracy of the results. We compared the behaviour (resolving power and artefactual resolution) of ordered parsimony, unordered parsimony and three-item analysis (3ta) in phylogenetic inference, as recently done by Grand et al. (2013).

Several methods are available to compare unrooted trees (Robinson and Foulds, 1981; Estabrook et al., 1985), but we prefer using rooted trees because they are the only classificatory structures which can convey unambiguous information about phylogenetic relationship (Rohlf, 1982). An unrooted tree conveys information, but it is ambiguous in the sense that it is compatible with several alternative relationships between taxa. Tree comparisons were performed with the ‘Inter-Tree Retention Index’ (ITRI) proposed by Grand et al. (2013) to measure the degree of congruence between two trees. The ITRI is based on the proportion of relationships (i.e. 3ts) between three OTUs, irrespective of their relationships with other OTUs that are common to two trees. An advantage of this method is that the computed tree-to-tree similarities are not symmetrical (i.e., the proportion of relationships is either the proportion for one tree or for the other). It is thus possible to discriminate between resolving power (power to find correct relationships) and artefactual resolution (incorrect relationships), by comparing trees obtained from simulated data with the tree used to generate these data. ITRI is equivalent to the retention index (RI) (Farris, 1989; Archie, 1989), calculated as a proportion of 3ts (Kitching, 1998). The ITRI is defined as:

\[
X_{RI} = \frac{2n - (X + 2n - X))}{(2n - n)}
\]

Where X stands for the sum of fractional weighting (FW) of 3ts implied by a character compatible with a given tree; n is the sum of FW of all 3ts obtained from the character. In each pairwise tree comparison, strict consensus trees were used to summarize congruence between results of an individual analysis (on a single matrix of 100 characters). See Grand et al. (2013) for a more detailed explanation. A strict consensus tree was built whenever analysis of a matrix produced more than one tree.

Our results are presented without mention of the number of optimal trees because the resolving power and artefactual resolution are the direct expression of data congruence and of the ratio homology/homoplasy, and that the ITRI expresses it well. Generally, a strict consensus tree of a thousand trees will show lower resolving power and artefactual resolution than a consensus of ten trees, because of data ambivalence. But there is always the possibility that an analysis will result in one false tree versus another analysis with one hundred slightly better trees. Kearney (2002) argued that ‘resolution of relationships is obviously a goal of phylogenetic analysis’, but her results show that phylogenetic performance cannot be measured by the number of trees. To avoid confusion with artefactual resolution, we complete her sentence by stating that it is correct resolution which is a goal of phylogenetic analysis. We think it is a necessary adjunction to remove the ambiguity between resolution and number of trees. Correct resolution does not display a simple relationship with the number of optimal trees.

**Testing the results**

Each consensus tree summarizing a phylogenetic analysis is compared with the reference tree using the ITRI to assess resolving power and artefactual resolution yielded by each method, each set of outgroup branch length, and each topology. Resolving power is calculated as the proportion of FW for 3ts of the reference tree that are also present in the consensus tree, and can be understood as the proportion of ‘true information’ the analysis has retrieved. Artefactual resolution is calculated as the proportion of FW for 3ts of the consensus tree obtained from an analysis that is not present in the reference tree, and this represents artefactual resolution yielded by the analysis. We calculated means of ITRI (i.e., we got a mean value for resolving power artefactual resolution on each reference topology). The statistical significance of differences in resolving power and artefactual resolution of the various methods of analysis was tested with the Wilcoxon signed-rank test (for paired samples) to compare means of ITRI, because there was no sample with normal distribution (Shapiro-Wilk test) and homoscedasticity of variances (Fisher test). Differences in topologies and branch lengths were tested using a Mann-Whitney test, and linear regres-
sions were performed to test tendencies were the external/internal branch length ratio varies, as in trees D, E, F and G (Fig. 5). Because many comparisons were made, we use the false discovery rate procedure (Benjamini and Hochberg, 1995) to assess the statistical significance of the differences in performance.

**Results**

All of the 2,700 analyses with equal branch lengths yielded strict consensus trees that are at least partly resolved; they all carry phylogenetic information. For ordered parsimony resolving power has a mean of 99.8% (sd: 1.1%) on the fully pectinate tree (Fig. 5A), and 91.1% (sd: 8.8%; Fig. 6) on the equiprobable tree (Fig. 5C with blue outgroup branch length). Articulat resolution is respectively about 0.2% (sd: 0.7%), 17.9% (sd: 7.2%) and 9.9% (sd: 6.3%). Resolving power is always a little lower and articulat reduction greater in 3ta than in ordered parsimony (p < 0.0001; Table 1). Unordered parsimony has the lowest resolving power (p < 0.0001), followed by the equiprobable topology, and by the fully pectinate topology (Table 1). With the 1,800 analyses performed on the palaeontological tree (Fig. 5D) and whatever method used, resolving power decreases and articulat resolution increases with outgroup branch length (p < 0.0001; Table 2; Fig. 7). Ordered parsimony seems to be less sensitive than 3ta and unordered parsimony to polarization errors (reflecting outgroup branch lengths).

From the 2,400 analyses with different branch lengths (Fig. 5D-G), the differences between the resolving power and articulat resolution values calculated on trees from a same reference topology cannot be explained by a linear model; results do not even vary consistently with the ratio between internal and terminal branch lengths (SS).

The impact of outgroup branch lengths was generally non-significant when we used trees with contemporary ingroup taxa (Fig. 5A-C). In 54 tests based on these trees, only 4 rejected the null hypothesis that outgroup branch length has no impact on performance (p < 0.05; Table 1). All 4 significant results concern the trees built from the symmetrical reference topology. Differences in the other 50 comparisons on the same trees (Fig. 5A-C) do not show any tendency of increasing or decreasing resolving power or articulat resolution when the outgroup branch length increases. Tests on trees built from the equiprobable palaeontological reference tree (Fig. 5D-G) with a variable outgroup branch length show that the performance of all methods decreases when outgroup branch length increases (Fig. 7; Table 2). Results of ordered parsimony and 3ta are very similar, with unordered parsimony yielding much lower performance.

Finally, we have compared the performance for each method on two trees with the same topology but in which the taxa were extant (Fig. 5C) or extinct (Fig. 5D) with a variable outgroup branch length. Artefactual reduction or decreasing resolving power or articulat resolution when the outgroup branch length increases. Tests on trees built from the equiprobable palaeontological reference tree (Fig. 5D-G) with a variable outgroup branch length show that the performance of all methods decreases when outgroup branch length increases (Fig. 7; Table 2). Results of ordered parsimony and 3ta are very similar, with unordered parsimony yielding much lower performance.

Reversals in phylogenetics

A particularly controversial issue in cladistics concerns the treatment of reversals. Proponents of parsimony (Kluge, 1994; Farris et al., 1995; Farris, 1997; Farris and Kluge, 1998) and 3ta (De Laet and Smets, 1998; Siebert and Williams, 1998) have been deeply divided on this particular issue. In parsimony, a transformation approach to homology using the Wagner (Farris, 1970) and Fitch (Fitch, 1971) parsimony algorithms treats characters from the perspective of unordered character-transformation trees (Slowinski, 1993). Reversals provide information that can serve a clad support because they are evidence of secondary homology with the appropriate test of maximizing congruence. This maximization of congruence leads to search the pattern with the minimum of ad hoc hypotheses that are convergences and reversals. For Farris (2012), reversals can be inferred from order primary homologies but also from an analysis: More
<table>
<thead>
<tr>
<th>Method</th>
<th>Symmetric</th>
<th>Pectinate</th>
<th>Equiprobable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ordered</td>
<td>Ordered/3ia</td>
<td>Unordered/3ia</td>
</tr>
<tr>
<td>Ordered/Unordered</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1/4</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>1/2</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>1</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

Table 1. Statistical significance of differences in resolving power and artefactual resolution (calculated through mean ITRI) between methods, outgroup branch lengths, and topology of the reference tree. The probabilities were calculated using Wilcoxon signed-rank tests for the comparisons between methods and with a Mann-Whitney test for the outgroup branch length and the shape of the reference tree. Non-significant results with α=0.05 are shown in blue, and non-significant results after correction for multiple tests using the false discovery rate (Benjamini and Hochberg, 1995) are shown in red. Individual values can be found in S7. Abbreviations: b.l., branch length.
fundamentally, even if (as I have seen other authors suggest) Henning would have preferred to distinguish apomorphies from plesiomorphies before starting to construct the tree, he was obviously willing to revise assessments of plesiomorphy during tree construction, for Henning did in fact recognize reversals and apply them as synapomorphies. This is the most widespread point of view in cladistics, which prevails in systematic paleontology, and the only point of view represented in probabilistic methods, which prevail in molecular systematic, but not in Henning’s systematic sense (Müller and Reisz, 2006). Assumptions of 3ta are much less familiar. In 3ta, hierarchical hypotheses of homology, i.e. a nested set of character states, are submitted to a test of congruence. The test either accepts or rejects the relevance of the hypothesis. Convergence is one of the multiple explanations of reversals: our results present a quantification of the loss of resolving power and artefactual resolution in 3ta if true and informative reversals are present (i.e. if true reversals are simulated and ‘hidden’ into the same state as the plesiomorphy but which support a clades of the known tree). Thus, our results must be interpreted ac-
cordingly. Irreversible characters might yield different results and will be tackled in another study.

We take this opportunity to propose a nomenclatural clarification about reversals (based on the example in Fig. 3A) as secondary homology hypotheses; thus, this clarification is valid both for parsimony and for 3ta. First rounds of reversals are generally called ‘secondary losses’ (e.g. Carine and Scotland, 1999), when in fact, only the absence should be considered secondary and the loss in itself should be considered as an event that appeared for the first time (i.e., primary). Thus, a character state is primitively absent (primary absence; state 0 on Fig. 3). It can then appear; this is a primary appearance (of state 1), denoted +1 on Fig. 3A. It can be subsequently lost (-1, reversal to state 0, but identified as 0* on Fig. 3a, for greater clarity); this should be called a primary loss, which results in a secondary absence. After this, a secondary gain (+2) can lead to sec-
ondary presence (1* in Fig. 3A), and a secondary loss (-2) can lead to ternary absence (0** in Fig. 3, etc.).

**Uncertain polarization**

We failed to find significant results on the impact of outgroup branch length and uncertainty in polarization on the nontaxonomic trees (Fig. 5A-C), except on some trees built from the symmetrical reference topol-
ygy. For instance, the evolutionary hypothesis of classifying 9 characters as well (Müller and Reisz, 2006). Assumptions of 3ta are much less familiar. In 3ta, hierarchical hypotheses of homology, i.e. a nested set of character states, are submitted to a test of congruence. The test either accepts or rejects the relevance of the hypothesis. Convergence is one of the multiple explanations of re-
versals: our results present a quantification of the loss of resolving power and artefactual resolution in 3ta if true and informative reversals are present (i.e. if true reversals are simulated and ‘hidden’ into the same state as the plesiomorphy but which support a clades of the known tree). Thus, our results must be interpreted ac-
cordingly. Irreversible characters might yield different results and will be tackled in another study.

We take this opportunity to propose a nomenclatural clarification about reversals (based on the example in Fig. 3A) as secondary homology hypotheses; thus, this clarification is valid both for parsimony and for 3ta. First rounds of reversals are generally called ‘secondary losses’ (e.g. Carine and Scotland, 1999), when in fact, only the absence should be considered secondary and the loss in itself should be considered as an event that appeared for the first time (i.e., primary). Thus, a character state is primitively absent (primary absence; state 0 on Fig. 3). It can then appear; this is a primary appearance (of state 1), denoted +1 on Fig. 3A. It can be subsequently lost (-1, reversal to state 0, but identified as 0* on Fig. 3a, for greater clarity); this should be called a primary loss, which results in a secondary absence. After this, a secondary gain (+2) can lead to sec-
ondary presence (1* in Fig. 3A), and a secondary loss (-2) can lead to ternary absence (0** in Fig. 3, etc.).

**Uncertain polarization**

We failed to find significant results on the impact of outgroup branch length and uncertainty in polarization on the nontaxonomic trees (Fig. 5A-C), except on some trees built from the symmetrical reference topol-
ygy. For instance, the evolutionary hypothesis (generated after an initial analysis by inferring clade history on a tree) involving three conditions deduced a poste-
riori from an analysis: 0 (‘absent’), 1 (‘present’) and 0* (‘secondary absence’; scored the same in a matrix but interpreted differently from primitive absence on a tree) is interpreted differently under parsimony and 3ta. The secondary absence can be explicitly represented as an apomorphy in the primary homology hypothesis (0(1(0(0*)))), under parsimony. Another interpretation (3ta) consists in disregarding secondary absence as synapomorphic but to consider it as a particular case of absence (0(1(0*))). Here, neither the absence nor the re-
versal is considered as a state (neither plesiomorphy, nor apomorphy) because the absence is not a state in 3ta (in Fig. 3, 0*, 0** and 0*** are not considered in 3ta). Parsimony proponents favour the first option, which yields support for six clades in the phasmatodea phylogeny (Fig. 3A). To summarize, the first interpreta-
tion (parsimony) considers a loss as an homology and a synapomorphy (because it supports a clade), an homo-
ploy (because the primary hypothesis is falsified by the other characters), and the other plesiomorphy (as defined in the matrix), according to Brower and de Pinna (2014). The second interpretation consid-
ers a loss as uninformative: it is neither an homology nor a synapomorphy (because it supports no clade), it is not an homoplasy (because the primary hypothesis is in agreement with the distribution of the other charac-
ters) and it is not a plesiomorphy (because the absence is not a state in 3ta; it is the root, including all). 3ta proponents favour this interpretation: only one clade in the Phasmatodea phylogeny is supported, and the only synapomorphy is the homology reflecting the first ap-
pearance of wings. These two interpretations are thus in perfect opposition. Here we emphasize that Browni-
an motion is only coherent with the assumptions en-
tailed by the first interpretation: reversals (i.e., second-
ary absence) are treated as apomorphies in the primary homology hypotheses (as an order with parsimony, or as a hierarchy with 3ta). Our simulations produce in-
formative reversals under Brownian motion, which can be exploited only under a parsimony viewpoint of these reversals: our results present a quantification of the loss in resolving power and artefactual resolution in 3ta if true and informative reversals are present (i.e. if true reversals are simulated and ‘hidden’ into the same state as the plesiomorphy but which support a clades of the known tree). Thus, our results must be interpreted ac-

containing an intermediate number of taxa (in our trees, clades with the maximal 3ta content have ten taxa inside and eleven taxa outside). These imbalanced clades will impact results only slightly compared to balanced clades. This may explain partly why tree shape influences phylogenetic reconstruction.

Comparisons of results between an ultrametric tree (Fig. 5C) and a paleontological tree of the same topology (Fig. 5D) show that the latter features better resolving power and artefactual resolution. This was obtained through FTRI mean percentages from the tree shown in Fig. 5D (also see S8).

**Implications on simulation-based studies and evolutionary models**

Our results highlight the advantages of models under which the relationships between character states can be represented by an unrooted tree (all molecular models) over 3ta hierarchical coding, if reversals can be simulated as in our study. Our simulations under Brownian motion can be represented as in Fig. 8. Firstly, characters are generated under a priori assumptions (here, the evolutionary model represented in Fig. 8A and the reference phylogeny shown on Fig. 8B). In a second step, these characters are interpreted as primary homology hypotheses by discretization and (for 3ta) conversion into hierarchical structures. Our procedure for simulation of characters, and by extension the use of evolutionary models, reflects quantitative characters that display informative reversals.

In parsimony, states that can be hypothesized to form morphoclines should be ordered if there is evidence. In this case, reversals can be represented by primary homology hypotheses. Empirical studies suggest that Brownian motion is a reasonable model for several types of characters, such as body size (Laurin, 2004), bone microanatomy (Canoville and Laurin, 2010), etc. Some characters do not seem to follow a strictly Brownian motion model, but instead may follow a speciation model (in which change occurs in both daughter-lineages at or near speciation, but no anagenesis takes place), such as morphological shape data in ratites (Laurin et al., 2012), a punctuational model (similar to the speciational model, but change occurs only on one branch) or an Ornstein-Uhlenbeck model (Uhlenbeck and Ornstein, 1930; Felsenstein, 1988). Some of these alternative models can be obtained from an irreversible evolution model is used (which can be understood as hierarchic). Another interesting field of research consists in developing methods that better simulate Darwinian evolution applied to digital life forms, as in the software Avida (Adami and Brown, 1994) to understand better the behaviour of parsimony and 3ta.

**Conclusion**

Our simulated characters evolved gradually; they are continuous and have a unimodal distribution (S6). They were discretized because phenotypic data are often represented as a proportion of total tree depth (0, 1/10, 1/3, 1/2, 2/3, 1) on resolving power and artefactual resolution.
available in discrete form (in descriptions, for instance), and most available parsimony and all 3ta programs require this. Thus, these reflect the data naturally analyzed by systematists. In this case, we created 10 states per character to retain much of the original information while producing data that can be easily analyzed by almost any phylogenetic package (Lauren in and Germain, 2011), and to allow detection of the spurious relationships yielded by homoplasy in the data (Bardin et al., 2013). This procedure yielded many significant results (Table 1; Table 2). We showed that unordered parsimony performs far worse than ordered methods on such data, with a loss in resolving power between 15.4% and 29.3% compared to ordered parsimony and with between 4.7% and 15.6% more artefactual resolution (results always statistically significant). Thus, a significant decrease in performance is expected when characters are not ordered as morphoclines. This result highlights the information content of character ordering schemes (Wilkinson, 1992), and we infer that coding continuous characters as fully unordered significantly decreases resolving power and increases artefactual resolution in empirical datasets. We also quantified the differences in resolving power between ordered parsimony and 3ta, which differ in the way they handle reversals as evidence for the support of a clade (among others). These differences vary between 1.3% (non-significant) and 6.3% (Table 1) for the resolving power, and between 0.0% (non-significant) and 9.4% for the artefactual resolution. We also have quantified resolving power when polarization errors, produced by varying outgroup branch length, are introduced (i.e., when polarization is based on an outgroup including a variable proportion of plesiomorphies). Further work could be done to quantify the effect of incorrect state ordering schemes as incorrect morphocline assumptions. Recently, Grand et al. (2013) found that 3ta yielded significantly greater resolving power and more artefactual resolution than ordered parsimony. In our study, based on an extended set of matrices, taxa and characters, ordered parsimony yielded the greatest resolving power and the fewest artefactual resolutions. The use of exact hypothetical ancestors to root trees in Grand et al. (2013) may explain the contrasting results. Moreover, the present work yields a better understanding of the impact of evolutionary assumptions about character state order and reversals. Further simulations could be done, using different evolutionary models, such as irreversible characters, to see if they yield different results about the performance of parsimony (with ordered states or not) and 3ta.

Acknowledgements
We are grateful to Jérémy Tissier, Laetitia Carrive and Eduardo Ascarrunz for fruitful discussions. We also thank Jacques Ducasse who developed modules of the Lixith program (on which the 3ta analyses were run) according to our needs, and to Régine Vignès-Lebel who thought of more efficient algorithms to store and compare sets of 3s deduced from the 3ta characters. Comments by Mark Wilkinson and an anonymous referee improved this paper. This work was financially supported by the UMR 7207 CRCP.

References
Online supplementary information

S1. Vba script used to generate the individual Nexus text files from the columns of discretized simulated data in Excel obtained using S2 and 7 and compiled into S4 (the latter is processed using S1 to get individual Nexus text files).

S2. Spreadsheet using the vba script presented in S1 to produce Nexus files for parsimony analysis.

S3. Spreadsheet using the vba script presented in S1 to produce Nexus files for 3ta analysis.

S4. The 1800 PAUP nexus files used in our parsimony analyses. Each sheet presents 100 matrices corresponding to one tree. Each sheet is identified by two numbers; the first number refers to the topology (1, pectinate; 2, symmetrical; 3, equiprobable), and second, branch length (sorted from shortest to longest). Thus, sheet 2.3 presents the matrices obtained from the symmetrical tree (Fig. 5B) with the third set of branch lengths (1). For the third topology, there are 12 branch lengths, corresponding to three branch lengths associated with Fig. 5C, six branch lengths associated with Fig. 5D, and with Fig. 5E-G).

S5. Resolving power and artefactual resolution in trees of various branch lengths.

S6. Frequency distribution of character states, based on 10000 characters in 21 taxa, obtained by discretization of continuous data generated by Brownian motion on the equiprobable tree (Fig. 5C, with outgroup branch length set at full tree depth).

S7. Individual ITRIs of the 1800 matrices (each of which was analysed using the three methods, for 5400 ITRI values for resolving power, and as many for artefactual resolution) used to compute resolving power and artefactual resolutions.

S8. Impact of branch length and topology on resolving power and artefactual resolution. Each cell value represents ITRI means for 100 matrices.