Impact of errors on cladistic inference: simulation-based comparison between parsimony and three-taxon analysis

Valentin Rineau, René Zaragüeta I Bagils, Michel Laurin

Contributions to Zoology, 87 (1) 25-40 (2018)

Impact of errors on cladistic inference: simulation-based comparison between parsimony and three-taxon analysis

Valentin Rineau[1, 3], René Zaragüeta i Bagils[2], Michel Laurin[1]

[1] *CR2P, UMR 7207, CNRS/MNHN/UPMC, Sorbonne Universités, 43 rue Buffon, F-75231 Paris cedex 05, France*
[2] *ISyEB (Institut de Systématique, Evolution, Biodiversité), UMR 7205 CNRS/MNHN/UPMC-EPHE, Sorbonne Universités, UPMC Univ Paris 06; Laboratoire Informatique et Systématique, France*
[3] *E-mail: valentin.rineau@upmc.fr*

Keywords: cladistics, homoplasy, maximum parsimony, model-based simulations

Abstract

Simulation-based and experimental studies are crucial to produce factual arguments to solve theoretical and methodological debates in phylogenetics. However, despite the large number of works that tested the relative efficiency of phylogenetic methods with various evolutionary models, the capacity of methods to manage various sources of error and homoplasy has almost never been studied. By applying ordered and unordered methods to datasets with iterative addition of errors in the ordering scheme, we show that unordered coding in parsimony is not a more cautious option. A second debate concerns how to handle reversals, especially when they are regarded as possible synapomorphies. By comparing analyses of reversible and irreversible characters, we show empirically that three-taxon analysis (3ta) manages reversals better than parsimony. For Brownian motion data, we highlight that 3ta is also more efficient than parsimony in managing random errors, which might result from taphonomic problems or any homoplasy generating events that do not follow the dichotomy reversal/convergence, such as lateral gene transfer. We show parsimony to be more efficient with numerous character states (more than four), and 3ta to be more efficient with binary characters, both methods being equally efficient with four states per character. We finally compare methods using two empirical cases of known evolution.

Contents

Introduction 25
Material and methods 27
Simulated character sets 27
HIV character set 29
Copied manuscript character set 30
Results 30
Ordering and number of states 30
Ordering and errors 31
Irreversible characters 32
Uncorrelated errors 32
Empirical tests 33
Discussion 33
Effect of the number of states 33
Errors in ordering schemes 34
Effect of evolutionary reversals and convergences 35
Random errors, homoplasy and phylogenetic inference ... 36
Empirical studies 37
Conclusion 37
Acknowledgements 38
References 38

Introduction

Phylogenies serve as a basis for most macroevolutionary studies. The reliability of phylogenetic trees depends partly on the characters used to construct them. Characters are the core of systematics, and questions such as "what is a good character?" or "how to code characters to best capture their evolutionary information?" are of critical importance for phylogenetic inference. Even though most recently published phylogenies have relied mostly or exclusively on molecular data, the phenotypic data on which we concentrate here are regaining a vital importance in systematics because some of the most promising recent methods to date the Tree of Life require use of morphological data. These are tip-dating, also known as total-evidence-dating (Pyron, 2011; Ronquist *et al*., 2012) and the more recent total-evidence-dating under the fossilized birth-death process (Zhang *et al*., 2016). Phenotypic data are required in these cases because extinct taxa need to be placed in the phylogeny and for most of these molecular data are unavailable, given that the oldest uncontroversial DNA sequences are less than 1 million year old (Willerslev and Cooper, 2005; Prüfer *et al*., 2010), whereas metazoan body fossils that are sufficiently diagnostic to be easily placed in phylogenies date back to the Cambrian (*e.g*. Vinther *et al*., 2014), about 495–541 Ma (Gradstein *et al*., 2012). Moreover, it may not be a sound methodology to omit the knowledge that morphological research has accumulated over several centuries.

Among little-investigated factors that can influence the outcome of phylogenetic analyses are the various kinds of errors that can creep in at various stages

of the process. Among these, scoring errors are a central problem of phylogenetic methods. Errors can produce incongruence between a character and the phylogeny, *i.e.* homoplasy, and hence, change the most parsimonious cladogram. The sources of coding and scoring errors that can inflate the apparent extent of homoplasy are potentially unlimited; these include, among others, typing errors in the matrix or erroneous interpretation of features (e. g. erroneous DNA alignment, data degradation). However, homoplasy can also result from difficulties of interpretation of the evolutionary process, as in lateral gene transfer, loss of morphological or molecular features, or convergence. Errors can also creep into all other stages of the phylogenetic analysis, through problematic species delimitations, errors in Operational Taxonomic Units (OTU) specifications, and errors in ordering schemes. Such errors are certainly more common than generally believed. Errors in coding phenotypic characters are due, in part, to the subjective judgement and unique knowledge of the researcher. In molecular analyses of nucleotide sequences this problem can occur because of subjective judgement during alignment and through different interpretations of insertions and deletions. Finally, suboptimal coding can lead to information loss, for instance when a character is coded with too few character states, which fails to capture part of the phylogenetic information.

Beyond plain errors, differences between phylogenies can reflect character treatment, which differs fundamentally in Maximum Parsimony (MP; Farris, 1970) and Three-taxon analysis (3ta; Nelson and Platnick, 1991; Nelson and Ladiges, 1992). The differences in how they deal with homoplasy originate from the particular way hierarchical characters are managed. The congruence between characters in MP is calculated by usually scoring the minimal number of transformations from a state to another, *i.e.* in unordered MP, without specifying the putative synapomorphy. Transformation events from any state to any other state can arise, regardless of their status as plesiomorphic or apomorphic. For instance, a clade can be supported in MP by a reversal, whereas the hierarchical coding of 3ta fails to offer such a support. Under 3ta, a character coded as (0(1)), is decomposed and can only provide three-taxon statements (3ts) composed of (A:0(B:1,C:1)), where A, B and C are terminal taxa. A 3ts of (A:1(B:0,C:0)) from an initial hypothesis of (0(1)) is meaningless in 3ta and represents a rejection of the 3ts (Nelson and Platnick, 1991). The differences in character representation between MP and 3ta cause significant differences in the way homoplasy and error are managed, and finally on the resulting cladogram. MP with a fully unordered coding is used in many phylogenetic studies on morphological datasets, especially in paleontology (*e.g.*, Liu, 2016; Lu *et al.*, 2016). Model-based approaches such as Maximum Likelihood and Bayesian Inference are not treated in this paper because they are computationally much more demanding and would have necessitated reducing the number of taxa and characters, which would probably have decreased the statistical significance of our results. Such studies would need a different design that is better adapted to material (computing power) constraints. A study comparing MP and Bayesian Inference with adapted datasets is in preparation.

Simulation-based systematics can help choose an appropriate phylogenetic inference method or a specific coding scheme. This branch of systematics relies on comparisons of the results of analyses of datasets simulated on known phylogenies; typically, a few parameters vary to make the results more general. A first attempt to empirically understand the differential behaviour of MP and 3ta using this approach was made by Grand *et al.* (2013). Their study rested on characters simulated under Brownian motion, allowing them to compare three cladistic methods of character treatment. They showed for intrinsically ordered characters that unordered parsimony is the method with the worst efficiency. Between ordered MP and 3ta, 3ta provided the highest resolving power, but also the highest artefactual resolution. Rineau *et al.* (2015) performed more extensive simulations by comparing, on larger data sets in continuous Brownian motion, the impact of polarization, branch length and tree shape on phylogenetic performance. They showed that ordered MP always performed better than 3ta for characters with 10 states. The results were explained by the differences between MP and 3ta in their treatment of homoplasy and, more precisely, of reversals.

We propose here an empirical study of the impact of homoplasy and scoring errors in phylogenetic studies with the same tools of experimental systematics. In order to complement the previous studies that our team performed on phylogenetic methods (Grand *et al.*, 2013; Rineau *et al.*, 2015), we deal here with the following interrelated issues:

1. The differential efficiency of MP and 3ta to coding variations. Our first analyses were made with an unrealistic number of states (10 per character) to exaggerate differences in performance. Even if morphological variation may be considered

unlimited, morphological characters are often described using only two states. Molecular characters (DNA sequences) have four or five states (five when coding insertions and deletions). We can hypothesize that when the number of states is reduced to better reflect most empirical studies, 3ta loses less phylogenetic information than MP because of its hierarchical coding. Indeed, in the simulations of Grand *et al*. (2013), where the number of character states was lower (binary characters), 3ts had in all cases a higher resolving power than MP. We thus test character sets with two, four, six, eight and 10 states.

2. Following the coding of character states and their number, we test the resilience of methods to the addition of errors directly in the character coding, more precisely, in the ordering scheme. The aim of these experiments, by adding more and more errors to the ordering scheme, is to determine under which circumstances unordered parsimony becomes preferable to ordered parsimony and 3ta. Ordered MP and 3ta are supposed to become less efficient than unordered MP as the ordering scheme degrades.
3. Our first simulations in Rineau *et al*. (2015) used Brownian motion, and showed greater efficiency for MP than for 3ta. We here attempt to compare these results on Brownian motion, with generation of convergences and reversals, to simulations under a model of irreversible evolution with only convergences. A heated debate among systematists (Harvey, 1992; Nelson and Ladiges, 1996; Platnick *et al*., 1996; Farris and Kluge, 1998; Zaragüeta and Bourdon, 2007; Farris, 2012) focused on the theoretical justifications of how to deal with reversals arose. In MP, reversals are not treated in a special way; they are transitions between states like any other. In 3ta, only derived character states are relevant for clade support by clade ancestry. Reversals in 3ta may be interpreted on a tree but are never used to positively support a clade. Our simulations address this problem by providing empirical results with which to compare predictions based on the theory.
4. We also assess the impact of random scoring errors in the matrix. Previously, this has been done infrequently (if at all) because earlier studies in experimental systematics have focused on molecular data, for which sequencing errors have not been of great concern. However, numerous problems may arise that are not correlated to the phylogeny at all, such as contaminations or alignment problems. In comparative anatomy, errors of interpretation by systematists are the most crucial problem. No datasets are exempt from such difficulties. Thus, we generated random errors in the data, which we termed "uncorrelated homoplasy" because it is uncorrelated to the true evolutionary tree. To increase the uncorrelated homoplasy ratio, we generated errors randomly with respect to their value and position in the initial matrix. We tested the effect of uncorrelated homoplasy through comparisons with datasets lacking uncorrelated homoplasy (with perfectly fitting datasets without homoplasy) or with others showing correlated homoplasy (with reversals and convergences). We test our hypothesis that 3ta may be more efficient at extracting phylogenetic information when such random errors arise.
5. Model-based simulations are relevant to the extent that they mimic genuine biological data. To assess the performance of phylogenetic inference methods through another approach, we analyse two empirical datasets for which the true evolutionary history is known without error. We used these two empirical examples to compare with results obtained through the simple evolutionary models we used in our simulations. The two examples detailed below are the known HIV transmission history, and an experimentally generated phylogeny of manuscript texts copied and recopied with typos.

A synthesis of all these results allows us to compare the efficiency of MP (ordered or not) and 3ta in various situations that have not been examined before.

Material and methods

Simulated character sets

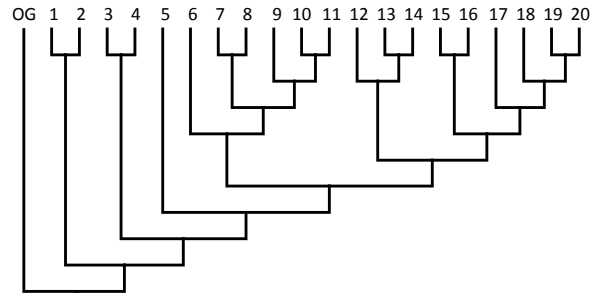

Figure 1. Ultrametric tree used in our simulations. The randomly generated phylogeny comprises an ingroup of 20 OTUs and a single outgroup (OG) OTU.

Reference tree. The data sets were simulated on a randomly generated topology by Mesquite (Maddison and Maddison, 2011). The tree comprised 20 OTUs (Figure 1), plus an outgroup to root the trees (to polarize character trees in 3ta). Our tree is ultrametric, and the information about topology is supplemented by 39 branch lengths for the ingroup and a branch length leading to the outgroup. The branch lengths are directly proportional to the expected variance of characters for the Brownian motion evolutionary model (Felsenstein, 1985). Branches allow a visual quantification of transformations, and can be equated to evolutionary time (OTUs are of the same geological age for our simulations given that our tree is ultrametric) under Brownian Motion.

Simulations and matrix coding. Continuous characters were simulated with Brownian motion on Mesquite. This simple model is very often used in evolutionary biology because it allows representation of characters with stochastic evolution, without selection pressure. This model is assumed in squared-change parsimony (Maddison, 1991), among other comparative methods. Simulated characters are inherently ordered because their continuous value is discretized into several states (in this case, of equal amplitude). The stochastic nature of Brownian motion also constrains the character states to display a unimodal distribution. Discretization is thus arbitrary, and methods such as gap coding (Mickevich and Johnson, 1976; Almeida and Bisby, 1984; Archie, 1985) are not especially helpful in this situation. First, 100 matrices of 100 characters were modelled. The 100*100 continuous characters were then discretized (into states with equal intervals). The discretization of each character into eight, six, four and two states allows us to test the efficiency of ordered MP, unordered MP, and 3ta with various numbers of states, for a total of 500 matrices; this is our first test.

The second test with Brownian motion addressed coding errors and, more precisely, errors in the ordering scheme. The simulation consists of 100 new matrices of 100 characters of 10 states (conditions used to test topology and polarization in Rineau *et al*., 2015), each with the coding of the true ordering reflecting the ideal situation for an analysis. Then, the true scheme was modified according to two alternative scenarios to introduce errors, as shown in Figure 2. The first scenario implied small errors: we produced 100 matrices with a single permutation between adjacent states for each character (Figure 2, scenario 1), and then added permutations iteratively to a maximum of five permutations per character. The second scenario differed in the severity of errors; in this case, the permutations were between states at opposite extremes of the transformation series, which should be far worse for ordered parsimony and 3ta (Figure 2, scenario 2). Each scenario of five coding iterations plus the perfect (initial) ordered scheme led to a total of 1100 matrices, analysed in ordered parsimony, unordered parsimony, and 3ta.

Another set of 100 matrices of 100 binary characters was generated with another evolutionary model, the asymmetrical Markov-k model. Parameters of the

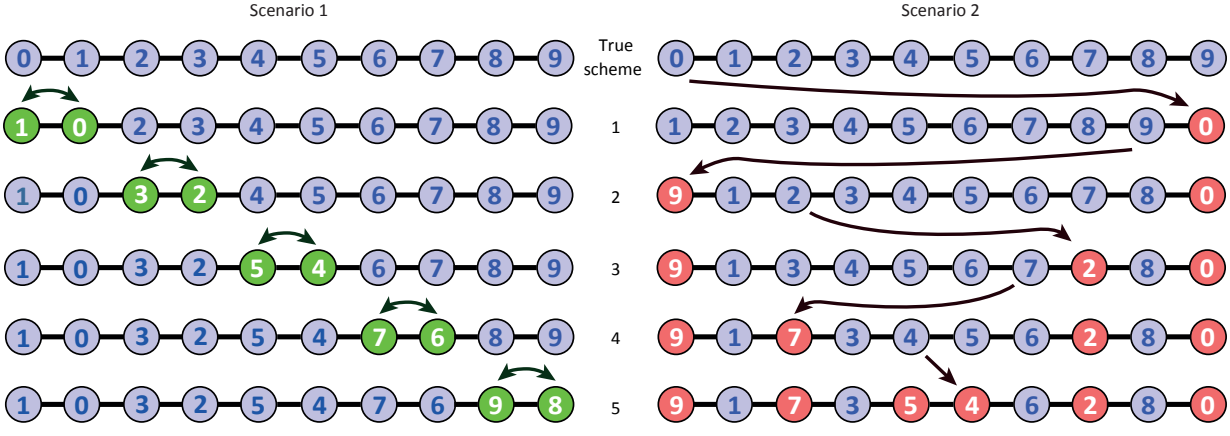

Figure 2. Two scenarios of errors in character ordering (in a character with 10 states) with five stages of degradation. Scenario 1 represents the case with slight permutation errors involving neighbours. Scenario 2 represents the worst case possible, with states displaced as far as possible from their neighbours. Green and red states represent permuted states and arrows represent permutations.

model were set to simulate irreversible characters by fixing the probability of going from states "0" to "1" at 0.1 and the probability of going from states "1" to "0" at 0.0 (forbidden).

These simulations tested only the effect of correlated, phylogeny-dependent errors and homoplasy. Finally, we took the 100 original matrices generated by Brownian motion with 100 characters and four states per character. We generated for each cell of each matrix a probability p=0.20 of change from the correct value generated under Brownian motion to a random value. This was done using the RAND function of Microsoft Excel 2013©. This allowed the generation of completely uncorrelated errors.

Tree searches. All 1700 matrices of 100 characters were analysed with both MP (ordered and/or unordered when characters had more than two states) and 3ta. The result of each analysis was represented by a strict consensus tree whenever more than one tree was found. All analyses were undertaken using PAUP* 4.0b10 (Swofford, 2003). In order to perform 3ta analyses, the matrices of 3ts with fractional weighting were generated using LisBeth 1.0 (Zaragüeta *et al*., 2012). All analyses performed on PAUP used the heuristic TBR algorithm with taxon addition and 50 random replicates after verification that this setting appeared to systematically find all or most of the shortest trees.

Tree comparisons. The optimal trees resulting from the analyses were then compared to the initial reference topology (Figure 1). To obtain percentages of accuracy for each analysis, the "Inter-Tree Retention Index" (ITRI) introduced by Grand *et al*., (2013) was computed. Several methods have been developed to compare phylogenetic trees. However, most use unrooted trees (Robinson and Foulds, 1981; Estabrook *et al*., 1985) to calculate distances between trees. Phylogenetic information, *i.e*. relationships between taxa, is nevertheless represented by a rooted tree. The ITRI measures a degree of congruence by using 3ts as elementary phylogenetic assertions. The method differs from that advocated by Critchlow (1996) by using fractional weighting, a correction to eliminate 3ts redundancy (Mickevich and Platnick, 1989; Nelson and Ladiges, 1992). The ITRI has also the advantage of differentiating true resolution (TR - a percentage of the true relationships: the relationships present in the reference tree that also occur in the obtained trees) from false resolution (FR - a percentage of the false relationships: the relationships that are not present in the reference tree that occur in the trees obtained by our searches). When necessary, we synthesized the general efficiency of an analysis using a standardized efficiency percentage, computed as (TR-FR+100)/2 (TR and FR are percentages; the efficiency TR-FR is comprised between -100 and 100; added to 100 and divided by 2, the efficiency is standardized in percentage). In such a measurement, an efficiency of 50% is uninformative (the number of correct clades equals the number of wrong clades). A lower efficiency indicates more artefactual than correct resolution. However, in these analyses, we are interested only in comparing the relative efficiency of ordered MP, unordered MP, and 3ta.

Test of the significance of the results. For each set of 100 matrices, two means are computed, one for TR and another for FR. The Wilcoxon test with signed ranks allows us to compare the ITRI computed from the same matrices but analysed by different methods, and with different kinds of discretization. We selected this non-parametric test because none of the samples follows a normal law (as shown by a Shapiro-Wilk test), and the variances were uneven (shown by a Fisher test). Linear regressions were computed to test the effect of the number of states and the number of errors on ordering schemes. The false discovery rate procedure (Benjamini and Hochberg, 1995) was used to discard false positives because of the high number of tests. All tests were computed on R 3.0.3 and XLSTAT Pro 2014.2.

HIV character set

The first empirical comparison between parsimony and 3ta was made using the HIV molecular sequences (population sequences from the *env* V3 and $p17^{gag}$ regions, known to be sufficiently variable for phylogenetic analyses) of Leitner *et al*. (1996), available from GenBank (Access numbers: from U68509.1 to U68521.1), for the ingroup. Two outgroup sequences were used, one African HIV dated from 1957 (GQ431830.1) to ensure that the virion was outside of the ingroup, whose last common ancestor is dated from 1981, and a SIV (Simian Immunodeficiency Virus) sequence from a macaque (AF041984.1), far more temporally distant from the ingroup. The sequences of 439 pb were manually aligned (Online Supplementary Information SOM 1). Insertion and deletion events were coded as binary characters. The two matrices include 14 taxa and 23 parsimony-informative characters (all unordered) with the HIV outgroup (29 with the SIV outgroup), and were analysed with parsimony on PAUP* 4.0b10 (SOM 2 and 3) and in 3ta on LisBeth 1.0 (SOM

4 and 5), both in branch and bound. Some sequences included polymorphisms. In 3ta, polymorphism produces a repetition of terminal taxa in the character-tree because a polymorphic terminal taxon is assigned to more than one terminal state. This repetition is caused by paralogy (Nelson, 1994; Fitch, 2000). Paralogy-free subtree analysis (Nelson and Ladiges, 1996; Zaragüeta *et al*., 2012), initially proposed for biogeographic paralogous taxa (created by the presence of taxa in more than one area), was used to deal with repeated terminal polymorphic taxa in the 3ta analyses.

Copied manuscript character set

The evolution of copied texts through time can be studied through theories and techniques very similar to cladistics (de Pinna *et al*., 2016). It is our second empirical comparison between parsimony and 3ta. The protocol of their generation is described in Spencer *et al*. (2004). A text used as an ancestral sequence was copied by volunteer scribes. Then, the previously handwritten copies were copied, and so on, thus generating a known phylogeny of the manuscripts. We aligned the words of the various text sequences by hand (SOM 6). The ancestral sequence contains 834 words (49 sentences). We coded a new character set revealing four types of possible events: (i) mutations (copying errors implying one or several letters of a single word; "ordinance" becoming "audience" is, for example, coded as a single event); (ii) permutations ("what, single-handed" becoming "single-handed, what"); (iii) fragmentation of a word into several ("gadflies" in "gad flies"); (iv) and insertions/deletions of words. The coding resulted in a parsimony and a 3ta matrix (respectively SOM 7 and 8). To limit search time (and keeping the matrix small enough to use a branch and bound algorithm for both methods), several OTU's were randomly suppressed to produce the matrices for parsimony (PAUP) and 3ta (Lisbeth). The matrix contains 13 taxa and is rooted on the known ancestral sequence. It contains 88 informative characters.

Results

Ordering and number of states

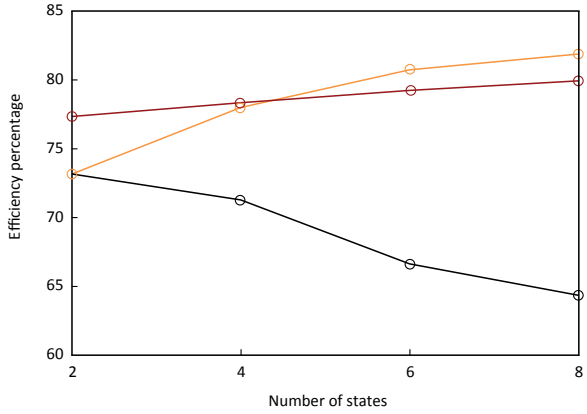

Figure 3. Impact of number of states in ordered, unordered MP, and 3ta on the efficiency of the methods. The difference in efficiency between methods was assessed by iterative recoding of the same matrices with two, four, six and eight states. Orange: ordered MP; red: 3ta; black: unordered MP.

Table 1. Impact of the number of states on MP and 3ta efficiency. Probabilities of the null hypothesis (no impact of number of states on efficiency) were tested through linear regressions.

number of states	Ordered MP			Unordered MP			3ta		
	TR	**FR**	**Efficiency**	**TR**	**FR**	**Efficiency**	**TR**	**FR**	**Efficiency**
2	69.89	23.55	73.17	69.90	23.55	73.17	81.90	27.20	77.35
4	76.34	20.33	78.00	68.22	25.63	71.29	83.14	26.43	78.35
6	81.91	20.42	80.74	64.47	31.20	66.63	84.02	25.52	79.25
8	84.28	20.51	81.88	58.60	29.90	64.35	84.27	24.39	79.94
p-value	**.**	**.**	**0.0387**	**.**	**.**	**0.0139**	**.**	**.**	**0.0032**
intercept	.	.	71.23	.	.	76.64	.	.	76.56
slope	.	.	1.44	.	.	-1.56	.	.	0.43
R^2	.	.	0.92	.	.	0.99	.	.	0.97

Table 2. Impact of errors in the ordering scheme on MP and 3ta efficiency, based on characters with 10 states. Probabilities of the null hypothesis (no impact) were tested through linear regressions. Two kinds of errors were tested: minor (scenario 1) and major (scenario 2) errors. See Figure 2 for explanations of both scenarios. Unordered parsimony for these data yields an efficiency of 69.0%.

	scenario 1		**scenario 2**	
number of errors	**MP**	**3ta**	**MP**	**3ta**
0	90.14	86.47	90.14	86.47
1	90.17	87.36	77.26	71.73
2	88.83	86.63	77.94	49.41
3	87.39	85.86	51.83	44.02
4	87.34	82.35	49.7	43.37
5	87.56	80.36	47.81	44.02
p-value	**0.0121**	**0.0203**	**0,.0048**	**0.0185**
intercept	90.2	88.15	88.67	78.12
slope	-0.65	-1.32	-9.15	-8.65

Efficiency increases linearly with the number of states (from two to four, six, eight and 10 states) for ordered MP and in 3ta. In contrast, efficiency in unordered MP decreases linearly with the number of states (Figure 3; Table 1). The efficiency of 3ta does not change exactly as MP. It is the most efficient method with binary matrices (77.3%, compared with 73.2% for MP). With four states, performance of both methods is similar (3ta: 78.3%, ordered MP: 78.0%) because MP efficiency increases quicker with additional states. Finally, MP becomes more efficient than 3ta when more than four character states are coded, although the difference between the methods is small (for six states, 3ta: 79.2% and ordered MP: 80.7%). The differences between 3ta and ordered MP can be analysed more precisely: in most cases the TR is higher in 3ta than in ordered MP (81.9% vs 69.8% for binary coding). Only with at least eight states do we see an identical TR of 84.3%. However, the FR is always slightly lower (between 3 and 6%) in ordered MP than in 3ta.

Ordering and errors

To test the impact of errors in the ordering scheme, the reference was unordered MP (efficiency: 69.0%). For the first set of errors (scenario 1, minor errors), ordered MP is always superior to unordered MP (about 18.3 to 21.2% more efficient), with only a very slight decrease of 2.6% from the smallest to the greatest number of errors (Figure 4). For the second set of errors (scenario 2, major errors), ordered MP becomes drastically less efficient than unordered MP between the second and the third error. The third to the fifth errors are dramatic in terms of efficiency, where ordered MP can lose a maximum of 42.3% efficiency. Linear regressions of these data show how efficiency decreases as the number of ordering errors increases (Table 2).

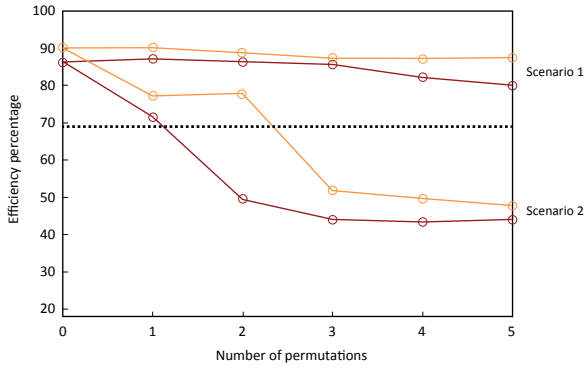

Figure 4. Impact of ordering errors shown in Figure 2 on MP efficiency. The difference in efficiency between methods was assessed by iterative recoding of the same matrices with the various ordering schemes with increasing the number of permutations. Scenario 1 shows MP's behaviour with slight errors; scenario 2 shows the worst possible errors in the ordering scheme. Orange: ordered MP; red: 3ta. Unordered MP is constant because the ordering scheme is not coded, being represented as a black dotted line.

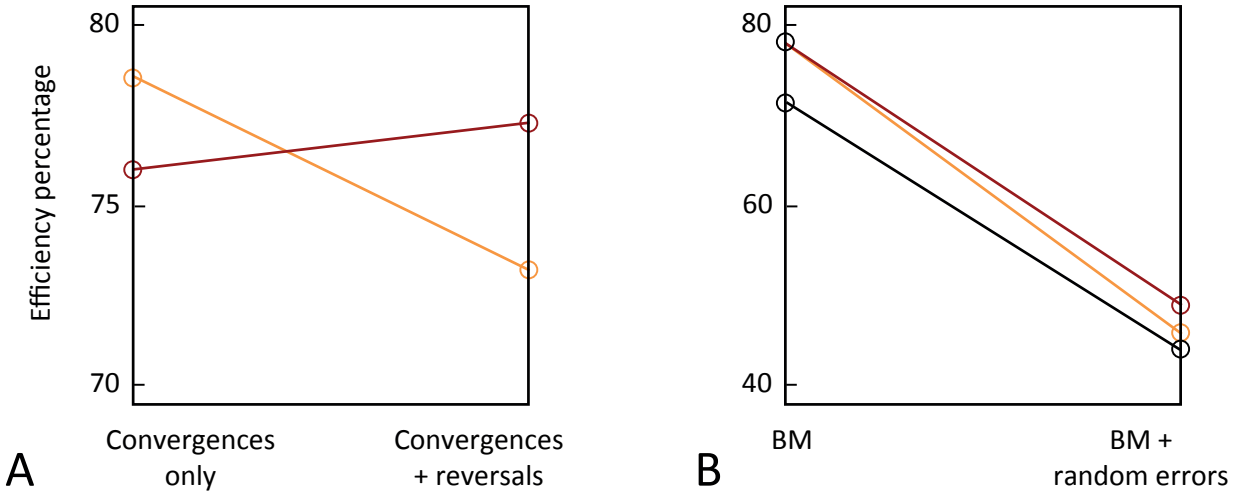

Figure 5. Impact of reversals (A) and randomly-generated errors (B) on phylogenetic reconstruction methods. A. Reversals. Left: the evolutionary model used is an irreversible model, generating only convergences, showing better performance of MP than 3ta (p-value: 0.02153). Right: a standard Brownian Motion model is used, generating both convergences and reversals; 3ta performs better than MP (p-value: 0.0004). B. Randomly generated errors. Left: standard Brownian motion model (no significant difference between methods; p-value: 0.4402). Right: same model with random errors added; 3ta performs better than MP (p-value: <0.0001). Orange: ordered MP; red: 3ta; black: unordered MP.

Table 3. Impact of reversals in the evolutionary model on MP and 3ta. An irreversible model (matrices modelled without reversals, with binary characters) is analysed with both methods and compared to Brownian motion with binary characters. A Wilcoxon test is used to test the significance of the difference of the results obtained with ordered MP and 3ta. Given that the characters used for this simulation are binary, there is no freedom in ordering scheme and hence, no difference between ordered and unordered MP.

	Reversal empirical test	
	BM	**irreversible model**
Ordered MP	73.2	78.6
3ta	77.3	76.0
p-value Ordered MP/3ta	**0.0004**	**0.0215**

Irreversible characters

As shown in Figure 5A and Table 3, MP efficiency is significantly higher under a model generating only convergent characters than under a Brownian motion model (BM) generating convergences and reversals (78.6% and 73.2%, respectively; p-value: < 0.0001). 3ta efficiency appears to be the same in both models (76 to 77.3%, p-value: 0.2572). The loss of efficiency of MP results from a 20.2% drop in TR (in BM compared to irreversible characters), partly compensated by a loss of 9.3% of false resolution (FR). Reversals in 3ta create the opposite effect: an increase of true resolution of 9.2% and an increase of 6.5% of FR. Unordered MP was not tested (separately from ordered MP) here because the data used for this test are binary.

Uncorrelated errors

The control is here on BM with four states (78.0% efficiency for ordered MP, 71.3% for unordered MP and 78.3% for 3ta; Table 4). Uncorrelated homoplasy (errors randomly added to the data matrix) added to simulations under BM decreases efficiency of all methods (Figure 5B). Unordered MP scores are the worst (44.4%) in this context. Efficiency of 3ta also decreases to 49.2%, but its performance remains significantly ($p < 0.0001$) superior to ordered MP (46%). We highlight another interesting result: 3ta always yields a single most parsimonious tree in these specific simulations, though not always fully resolved (in all of the 100 analyses), whereas MP yields an average of three most parsimonious trees per analysis.

Table 4. Impact of homoplasy by random errors on phylogenetic reconstruction on ordered MP, unordered MP and 3ta. Random errors are generated on matrices generated with Brownian motion on four states and compared to analyses of the same matrices without random error. A Wilcoxon test is used to test the significance between ordered MP and 3ta.

	Random error test	
	BM	**BM with random error**
Ordered MP	78.0	46.0
Unordered MP	71.3	44.4
3ta	78.3	49.2
p-value Ordered MP/3ta	**0.7002**	**> 0.0001**

Empirical tests

All characters are unordered in the analyses of empirical datasets. In the HIV example, 3ta is more efficient than unordered MP with both outgroups (Table 5). 3ta efficiency is 96.3% for the HIV outgroup and at 57.6% for the SIV outgroup. It is between 17.6% and 5.0% more efficient than MP, which shows an efficiency of 79.1% with the HIV dataset, and 52.6% with SIV. With the recoded dataset from Spencer ***et al.*** (2004), the results are more similar between methods (Table 6) and the dataset seems almost equally well handled by both methods. MP is slightly more efficient than 3ta with a higher TR and a lower FR, and an efficiency of 92.7% versus 91.7% for 3ta.

Discussion

Effect of the number of states

When comparing ordered and unordered MP using data generated through BM (Figure 3), our results show a linear correlation between the number of states and the efficiency of all methods, with contrasting tendencies (Table 1). The efficiency of ordered MP (and 3ta, to a lesser extent) increases with the number of states (when the ordering scheme is correct). This is expected because the phylogenetic information content of character states increases linearly with the number of states if the latter are ordered (*i.e.* a binary character can support only one clade, but a ternary ordered character will support two nested clades). However, this is not true for unordered states because in the worst-case scenario in which the number of states equals

Table 5. Empirical test of cladistic analysis by unordered MP and 3ta of HIV virion. The analysis is made with two alternative outgroups, a close HIV and a much more remote SIV group. Results present the number of trees for each analysis, the number of steps for MP (the concept being meaningless in 3ta), the retention index (RI), and the efficiency composed of TR and FR.

		VIH phylogeny				
		Number of trees	**RI**	**TR**	**FR**	**Efficiency**
MP	VIH	49 (95 steps)	0.69	58.2	0	79.1
	VIS	156 (217 steps)	0.64	38.6	33.5	52,5
3ta	VIH	1	0.74	97.1	4.6	96,2
	VIS	1	0.85	54.9	39.9	57.5

Table 6. Empirical test of cladistic analysis by MP and 3ta of copied manuscripts. Results present the number of trees for each analysis, the number of steps for MP, the retention index (RI) of the characters on the obtained tree, and the efficiency composed of TR and FR.

	Manuscripts phylogeny				
	Number of trees	**RI**	**TR**	**FR**	**Efficiency**
MP	2 (185 steps)	0.68	100	14.6	92.7
3ta	1	0.76	99.8	16.5	91,6

the number of OTUs, an unordered character has no phylogenetic information content, whereas an ordered character has a maximal information content (it would specify a fully-resolved tree, if not contradicted by other characters). Thus, splitting characters into many states should not be undertaken when relationships between character states are poorly constrained. On the contrary, when relationships between states appear clear, as for linear morphoclines, maximizing the number of ordered character states will maximize the general efficiency of MP to retrieve the correct tree.

Previous results on the comparative efficiency of the two ordered methods, MP and 3ta, have been contradictory (Grand *et al*., 2013; Rineau *et al*., 2015). Our simulations show a relationship between the number of states and the efficiency of both methods, which explains the contradictory results in previous studies. MP is more efficient than 3ta when the number of states per character is higher than four (Figure 3); this was the case for the study of Rineau *et al*. (2015), which simulated characters with 10 states each. With four character states, as in most molecular studies, ordered MP and 3ta have the same ability to find the true phylogeny (p-value: 0.4402). However, 3ta performs better than MP with binary characters, which is the most common situation in paleontological studies, with 4.2% more efficiency (p-value: 0.0004). This explains why Grand *et al*. (2013) found more favourable results for 3ta, because most of their characters contained three or fewer states. In simulations with binary characters, the artefactual resolution is always significantly higher in 3ta than in MP (3.6 to 6.1% higher; Table 1). This difference in resolution reflects the fact that the number of most parsimonious trees is always lower in 3ta, leading to a more resolved strict consensus. Nevertheless, 3ta retains more correct phylogenetic information than MP with decreasing number of states. The true resolution at eight states is equivalent; 2.1% better in 3ta with six states, 6.8% better with four states and 12% better with binary characters. The true resolution decreases in MP with decreasing number of states mostly because resolution decreases. To sum up, the 3ta seems to be able to deal with information of binary or ternary characters, while MP might better capture the information contained in the ordering scheme of characters with more than four states.

Errors in ordering schemes

The objective of the second set of simulations was to compare the performance of ordered and unordered methods when the ordering scheme contains errors (Figure 4; Table 2). Adding errors to the coding scheme of simulations under Brownian motion decreases the efficiency of both ordered MP and 3ta, as expected, but not to the same extent. The comparison with unordered MP shows that ordering characters strongly constrains the hypothesis. To investigate precisely the change in efficiency in the relationship with the number of errors, the analyses were performed on characters with 10 states. 3ta under BM with more than four states is always less efficient than MP, so ordered and unordered MP are emphasized in this discussion. Errors of ordering can vary in many ways. In the first scenario, when errors are minor (Figure 2), ordered MP has always a strong advantage, as compared to unordered MP (approximately 20% more efficient). In the second and worst-case scenario, ordered MP efficiency decreases dramatically when three or more inversions are present (six or more states are wrongly placed in the scheme), and becomes less powerful than unordered MP. Thus, even with one or two dramatic errors (inversions) in the coding scheme, ordered MP remains more efficient than unordered MP. More than half of the character-state positions (one of the worst-case scenarios when ordering states) need to be mistaken to make ordering disadvantageous with MP. With 3ta, results are always slightly worse than for ordered MP, with the advantage of the ordering scheme being lost, in the case of severe errors, with two (rather than three) or more inversions between states (Table 2). This poor efficiency of 3ta in this case might be partly due to the high number of character states used here (10), a behaviour already highlighted by Rineau *et al*. (2015). This result was expected because in 3ta hypotheses of homology are phylogenetic arguments addressed to support a phylogenetic hypothesis. If the arguments are massively inconsistent, one of the main theoretical premises of the method is violated.

Thus, several subtle errors in the ordering scheme are not enough to suppress the advantage of ordered analyses. For this advantage to be lost, the ordering scheme must be severely flawed. An example would be using a similarity criterion where in reality genetic determinants cause a discontinuous evolution that does not yield clines, such as merosity in flowers (the fact for floral pieces to be organized in multiples), in which case going from three axes to six is probably easier in certain clades than going from three to four axes (Decraene and Smets 1994).

Effect of evolutionary reversals and convergences

The aim of this set of simulations was to test the efficiency of cladistic methods on phylogenies constructed with a model forbidding reversals (but convergences were allowed). These results were then compared to those obtained by analysing datasets constructed with a model that allows reversals. We aimed to test their differential performance because proponents of MP (Harvey, 1992; Kluge, 1994; Farris *et al*., 1995; Farris and Kluge, 1998; Farris, 2012) and 3ta (Nelson, 1996; Nelson and Ladiges, 1996; Platnick *et al*., 1996; Siebert and Williams, 1998; Nelson *et al*., 2003; Williams and Ebach, 2005; Zaragüeta and Bourdon, 2007) have crystallized heated debates on how methods must manage reversals on a theoretical basis. The handling of reversals in 3ta and MP reflects great differences in how characters are conceptualized in these approaches. Taxic homology used in 3ta differs in the treatment of plesiomorphies from the transformational homology used in MP (Carine and Scotland 1999; Scotland 2000; Brower 2014; Farris 2014). For 3ta proponents, systematists propose putative hypotheses of synapomorphy. The root of a character may or may not be a state; in the former case, it is a plesiomorphy (Nelson, 1994). At a defined taxonomic level of an analysis, 3ta may consider a loss as uninformative because it is neither a homology nor a synapomorphy (Rineau *et al*., 2015). However, the most common way to deal with homology is that used in MP, in the context of transformational homology. If 0 is a plesiomorphy and 1 an apomorphy, the transformation from 0 to 1 supports a clade. However, MP also supports a clade if another transformation appears in the tree from 1 to 0. Moreover, MP may support both transformations at the same time, for different clades. Thus, evolution of the characters as in matrices (MP) or their nested status (3ta) need to be carefully assessed under both approaches.

One of the main arguments used against 3ta was its inability to recognize 'reversals' (Farris *et al*., 1995; De Laet and Smets, 1998; Farris and Kluge, 1998; Farris *et al*., 2001; Farris, 2010; Farris 2012). Critics have called reversals plesiomorphic character-states that are considered synapomorphies at less inclusive nodes through the minimization of the Manhattan distance implemented in MP. Simulations with irreversible evolution were made to provide new empirical data relevant to the theoretical discussions on reversals and 3ta. Their purpose was to verify how 3ta, which does not modify proposed hypotheses of homology so that they fit the tree, compares in performance with MP. If 3ta is unable to manage reversals in phylogenetic reconstructions, the method should have a drastically lower efficiency in the presence of reversals than in irreversible evolution. Our results (Figure 5A; Table 3) falsify this assumption because MP efficiency collapses when adding reversals in the evolutionary model, whereas 3ta efficiency remains stable (p-value: 0.2572 between 3ta with and without reversals in the evolutionary model). This result might be linked with the fact that in 3ta reversals increase the number of incompatible 3ts, whereas under MP reversals increase the number of evolutionary steps. Supernumerary steps may distort the entire character, whereas 3ts divide the character into minimal parts that can be independently retained or rejected. Our results show for the first time that the main criticism against 3ta may be empirically unfounded; in the absence of reversals, performance of 3ta should have been best, but that is not what we observed.

Conversely, with an irreversible evolutionary model, MP performs statistically better than 3ta (2.6% better, p-value: 0.0215). The explanation for this phenomenon is probably the divergence of character representation between steps (instance of transformations between states) and 3ts (statement of kinship relationship based on a derived state). Take the example of a homoplastic character state that support two clades in a tree that implies a convergence. These two clades in MP are supported by independent steps. It does not matter where these convergent transitions supporting both clades are situated on the topology. In 3ta, the closer the clades are to each other, the higher the number of 3ts will support the tree. The example on Figure 6 illustrates the difference between MP and 3ta. Regarding only the character X, both trees A and B are equivalent under MP, as the character distribution implies two steps on both topologies. The hierarchical 3ta character (*a*, *b*, *d*, *e*, (*c*, *f*)) is completely incompatible with the phylogeny A ((*a*, (*b*, *c*))), (*d*, (*e*, *f*))). On this character, no 3ts can be retained. Regarding only that character, this phylogeny is less optimal than the second phylogeny B (*a*, ((*b*, *c*), (*d*, (*e*, *f*)))), which is compatible with one 3ts *a*, (*c*, *f*), because of the proximity between the taxa that bear the derived state. The criterion for choosing a phylogeny in MP is to minimize the number of steps. In 3ta, it is to choose the most congruent phylogeny, the one which agrees with the maximum number of phylogenetic hypotheses represented by 3ts. While the supernumerary transformation steps are the same in

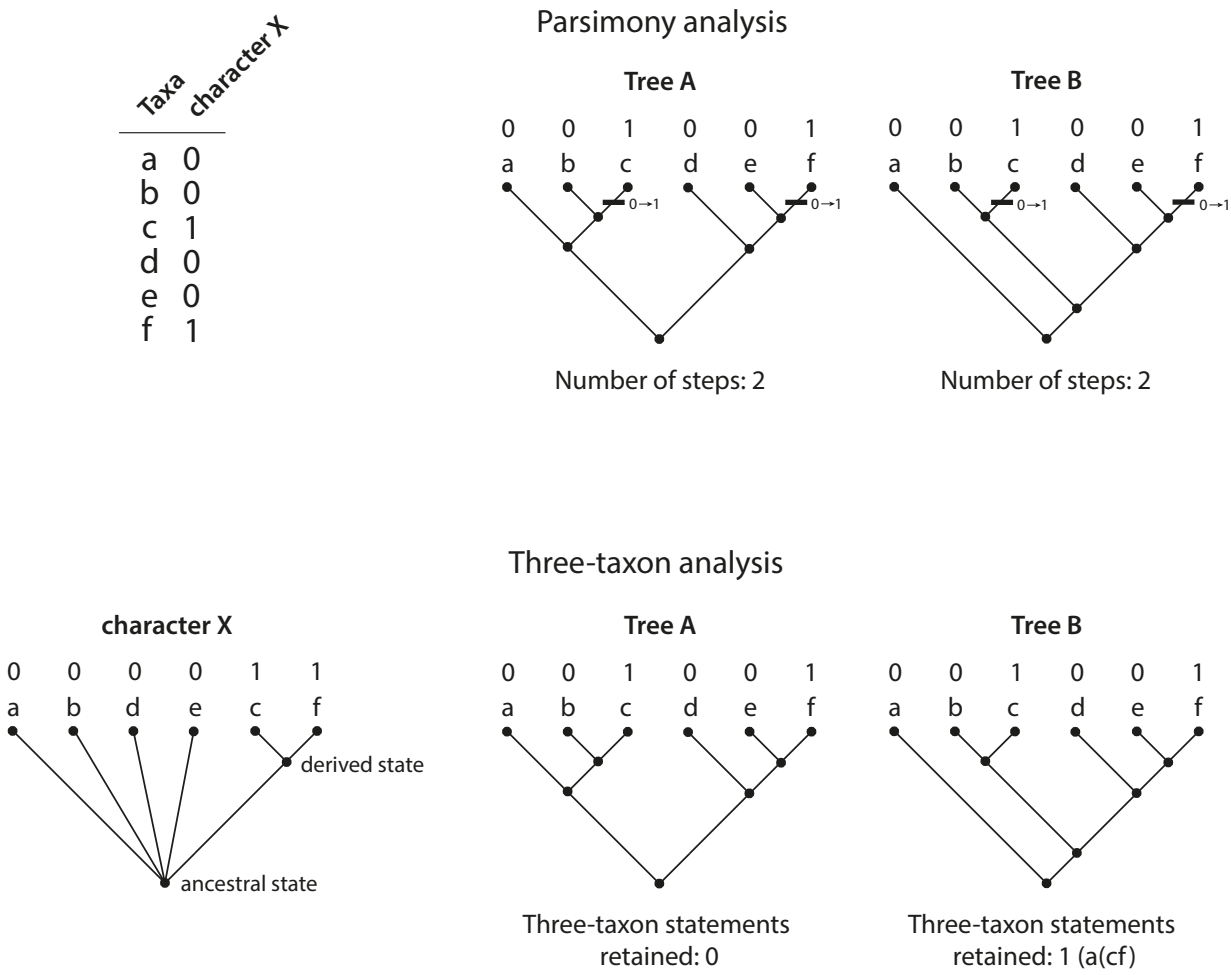

Figure. 6 Example of differential behaviours between MP and 3ta. The example shows two phylogenies, A and B, and a character X. Character X is represented in a matrix in MP and through a hierarchical character (a, b, d, e, (c, f), with (c, f) as a derived character state. In MP, the optimization of both trees always leads to two steps, regarding only character X; they are both suboptimal. In 3ta, the optimal tree is chosen as the structure retaining the maximum number of hypotheses (i.e. 3ts resulting from the character decomposition). Tree A is completely incompatible with character X, as no 3ts are retained. Regarding only character X, tree B is a better phylogeny as it retains one 3ts from the initial character. In 3ta, 3ts tend to 'attract' nodes that independently acquire a derived state, whereas this phenomenon does not occur in MP.

both topologies under MP, there is a 'node attraction' between clades sharing homoplastic convergent states in 3ta. This may reflect the difficulty of 3ta to deal with convergent features, especially when they are far from each other on the tree.

Random errors, homoplasy and phylogenetic inference

The purpose of our final set of simulations was to propose an alternative to the standard experimental simulations usually performed to compare phylogenetic methods (Wright and Hillis 2014; Goloboff *et al*., 2017). In standard evolutionary models, homoplasy is always of evolutionary significance. In evolutionary models, errors are always considered only as convergences and/or reversals. In these final simulations we added random errors to matrices produced by BM. These random errors simulate errors unrelated to evolution, such as human errors. These include errors of interpretation of the DNA alignment, errors of interpretation of structure due to taphonomical problems (Sansom *et al*., 2011), erroneous character coding, or simply mistakes when scoring cells in a matrix (such as ambiguity of the sequencing at a particular position,

or a measurement error on a structure). Problems of homoplasy generating events that do not follow the dichotomy reversal/convergence, such as lateral gene transfer (generating homoplasy when forced on a hierarchical topology), can be ranked under these "random errors" (a phylogenetic network becomes partially random from a hierarchical point of view). The differences between errors in the matrix and genuine homoplasy – phylogenetically structured homoplasy such as convergences and reversals – have been little investigated, but this topic is relevant to compare methods in different situations. For example, the interpretation of characters in DNA sequences must be a little less sensitive than comparative anatomy to interpretative human errors, although other sources of error, such as alignment ambiguity or sequencing errors, may be common. The impact of such errors on phylogenetic analyses needs to be minimized by using methods that do not over-interpret the dataset. Differences in efficiency were detected following addition of interpretative errors to the dataset (both with BM and with perfectly congruent datasets without convergences or reversals). Even though MP and 3ta lose efficiency when 20% of random errors are added, 3ta is 3.2% more efficient than ordered MP in the presence of such errors to reconstruct the true phylogeny (Figure 5B; Table 4).

The fact that reversals are never used to support clades in 3ta may limit over-interpretation of their evolutionary significance. 3ta integrates into its definition the possibility of error, defining homology as only *"the relationship among parts of organisms that provides evidence for common ancestry"* (de Pinna 1991; Brower and de Pinna 2012, 2014). MP, in contrast, will maximize phylogenetic information through minimization of supernumerary steps. Using this method means assuming that if a hypothesized synapomorphy is rejected, the method might generate several other synapomorphies (implying supernumerary evolutionary steps) to fit inconsistent character-states, which may lead to an over-interpretation of the original characters and an increase of false resolution.

Homoplasy generated by interpretative errors and all types of human errors may be crucial in phylogenetic reconstructions given that such errors may not be uncommon in empirical datasets (*e.g.* Marjanovic and Laurin 2008, 2009). The comparison of phylogenetic methods should address their robustness to such errors in the data. We highlight for the first time through simulations the capacity of 3ta to handle this issue significantly better than MP. In systematics based on phenotypic data, human errors are added to other sources of homoplasy such as convergence and reversals, and should be better taken into account in future studies.

Empirical studies

The first set of analyses on HIV virion supports our conclusions drawn from the simulations. The analyses using the nearest HIV outgroup show 3ta to be much more efficient than MP (Table 5). The main reason for the low performance of MP is certainly due to the high number of optimal trees (49 MPT), as shown in all our previous simulations. However, we note that the true tree is included neither among the 49 MP trees, nor concerns the single 3ta tree. With a farther outgroup (SIV), both methods decrease in efficiency, but 3ta remains more efficient than MP. However, 3ta efficiency seems to decrease more quickly with a more distant outgroup (38.7% decrease against 26.5% for MP). This sharp decrease may reflect difficulties of 3ta in dealing with distant outgroups with long branches, as shown by Rineau *et al*. (2015). These results may also reflect the different ways in which MP and 3ta use outgroups. In 3ta rationale, outgroup rooting is better used to root trees of character states. This strategy is more sensitive to outgroup choice than the MP handling because it assumes the outgroup to bear only plesiomorphic states, an unrealistic supposition. Finally, we show with 3ta analyses that a small number of informative characters (23) is sufficient to infer correctly a phylogeny when the polarization is reliable (outgroup criterion here).

The dataset on manuscript characters favours MP, which reaches its highest efficiency in this study (92.7%). 3ta shows slightly less efficiency (91.6%), because the true tree contains polytomies (Spencer *et al*., 2004). Given that our results are synthetized by a strict consensus, the fact that MP retains more optimal trees than 3ta is advantageous here. Given that 3ta yields a less optimal, fully resolved tree, polytomies (uncertainties) are not represented, pushing up the score of false resolutions.

Conclusion

We used simulations here to propose new empirical arguments to fuel methodological discussions on character coding and phylogenetic methods and to

develop new and more realistic evolutionary models. Debates between proponents of MP and of 3ta, and also on character coding and ordering, have been mainly based on theoretical considerations. Model-based studies allow comparison of phylogenetic analyses using simulations of characters following models on reference trees. They provide a way to compare analyses by varying a single parameter. Empirical examples can also be found in the literature of known genealogies (Leitner *et al*., 1996; Spencer *et al*., 2004), while our study has used both approaches.

The effect of the number of states per character is highlighted by our simulations (Figure 3). Evidently, more character states in unordered MP lead to a greater number of compatible topologies and to a quick decrease of efficiency, when the underlying evolutionary model logically leads to ordered states. The ordering of character states appears to be essential to retrieve the true tree. Both 3ta and MP with ordered characters see their efficiency increase with the number of character states. However, the two methods do not perform optimally under the same conditions. Their ability to retrieve the phylogeny is roughly equal with four states per character, MP being more efficient with more than four states, whereas 3ta is more efficient with fewer than four states. Because multistate characters with five or more states are rare in matrices, 3ta seems more efficient for most phenotypic datasets. The simulations referred to in the previous paragraph were made with ordering schemes that were known and error-free. We performed additional simulations with controlled errors to determine under which conditions (with known amount of ordering error) ordering characters is useful. The results show that analyses using ordered datasets should be discouraged only when several severe ordering errors are introduced. Unordered MP performs correctly only when we can suspect several drastically different but equally plausible ways to order characters. Otherwise, ordering following morphoclines (for example) seems preferable.

Irreversible simulations were used to offer empirical arguments to fuel the debate on reversal management in phylogenetics, which so far has relied mostly on theoretical considerations. 3ta using taxic homology seems to perform better than MP and transformational homology in the presence of reversals. MP may over-interpret data by supporting clades with reversals (Brower and de Pinna 2012), leading to artificially supported false clades and a drastic increase of the number of most parsimonious trees as compared to 3ta. Alternatively, 3ta may succeed better than MP at sorting primitive absence from losses.

Our final simulations assessed the impact of random errors (not correlated with the phylogeny) that might be explained by errors that have nothing to do with evolution, such as mistakes, interpretation errors, alignment problems and taphonomical losses (Sansom *et al*., 2011). In our simulations, the first to include this new parameter (as far as we know), 3ta is also more efficient than MP, possibly because of the over-interpretation of inconsistent data by MP. This new source of error should be better integrated in future model-based simulations for experimental systematics because it can be a parameter of critical importance (*e.g*., Marjanovic and Laurin 2008, 2009). Recent simulation-based studies of phylogenetic methods typically include probabilistic methods such as Bayesian inference and maximum likelihood (Goloboff *et al*., 2017 and references therein). All probabilistic methods implement transformational homology. Consequently, they might react to human error in a way similar to MP, but this need to be assessed.

Lastly, empirical studies put into perspective the fact that the ability of 3ta to retain fewer optimal trees may be advantageous or not, depending on the real tree. The greater resolution provided by 3ta proved useful to deal with the HIV dataset, but disadvantageous with the manuscript dataset, in which MP performed slightly better.

Acknowledgements

We thank three anonymous reviewers for their comments. We are grateful to Veronique Barriel for advice on coding molecular characters in parsimony, and to Matthew Spencer for all documents relating to the cladistic analysis of the Parzifal manuscripts. This work was funded by the Centre de Recherche sur la Paléontologie et les Paléoenvironnements, (UMR 7207 - CNRS, MNHN, UPMC).

References

Almeida MT, Bisby FA. 1984. Simple method for establishing taxonomic characters from measurement data. *Taxon* 33: 405–409.

Archie JW. 1985. Methods for Coding Variable Morphological Features for Numerical Taxonomic Analysis. *Systematic Biology* 34: 326–345.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society* 57: 289–300.

Brower AVZ. 2014. Transformational and taxic homology revisited. *Cladistics* 31: 197-201.

Brower AVZ, de Pinna MCC. 2012. Homology and errors. *Cladistics* 28: 529–538.

Brower AVZ, de Pinna MCC. 2014. About nothing. *Cladistics* 30: 330–336.

Carine MA, Scotland RW. 1999. Taxic and transformational homology: different ways of seeing. *Cladistics* 15: 121–129.

Critchlow DE, Pearl DK, Qian C. 1996. The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology* 45: 323–334.

De Laet J, Smets E. 1998. On the three-taxon approach to parsimony analysis. *Cladistics* 14: 363–381.

Decraene LR, Smets EF. 1994. Merosity in flowers: definition, origin, and taxonomic significance. *Plant Systematics and Evolution* 191: 83–104.

Estabrook GF, McMorris FR, Meacham CA. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology* 34: 193–200.

Farris JS. 1970. Methods for Computing Wagner Trees. *Systematic Zoology* 19: 83-92.

Farris JS. 2010. Systematic foundering. *Cladistics* 26: 1-15.

Farris JS. 2012. 3ta Sleeps with the fishes: Book review. *Cladistics* 28: 422–436.

Farris JS. 2014. 'Taxic homology' is neither. *Cladistics* 30: 113–115.

Farris JS, Källersjö, M, Albert VA, Allard M, Anderberg A, Bowditch B, Bult C, Carpenter JM, Crowe TM, Laet J, Fitzhugh K, Frost D, Goloboff P, Humphries CJ, Jondelius U, Judd D, Karis PO, Lipscomb D, Luckow M, Mindell D, Muona J, Nixon K, Presch W, Seberg O, Siddall ME, Struwe L, Tehler A, Wenzel J, Wheeler, Q, 1995. Explanation. *Cladistics*, 11, 211–218.

Farris JS, Kluge AG. 1998. A/The Brief History of Three-Taxon Analysis. *Cladistics* 14: 349–362.

Farris JS, Kluge AG, De Laet J. 2001. Taxic Revisions. *Cladistics* 17: 79-103.

Felsenstein J. 1985. Phylogenies and the comparative method. *The American naturalist* 125: 1–15.

Fitch WM. 2000. Homology: a personal view on some of the problems. *Trends in genetics* 16: 227-231.

Goloboff PA, Torres A, Arias JS. 2017. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics* (Early View), doi: 10.1111/cla.12205.

Gradstein FM, Ogg JG, Schmitz M, Ogg G. 2012. The geologic time scale 2012. Amsterdam: Elsevier.

Grand A, Corvez A, Duque Velez LM, *et al*. 2013. Phylogenetic inference using discrete characters: performance of ordered and unordered parsimony and of three-item statements. *Biological Journal of the Linnean Society* 110: 914–930.

Harvey AW. 1992. Three-taxon statements: more precisely, an abuse of parsimony? *Cladistics* 8: 345–354.

Kluge A. 1994. Moving targets and shell games. *Cladistics* 10: 403–413.

Laurin M. 2010. How Vertebrates Left the Water. Berkeley: University of California Press.

Leitner T, Escanilla D, Franzén C, *et al*. 1996. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences* 93: 10864–10869.

Liu J. 2016. *Yuanansuchus maopingchangensis* sp. nov., the second capitosauroid temnospondyl from the Middle Triassic Badong Formation of Yuanan, Hubei, China. *PeerJ* 4: e1903.

Lu J, Zhu M, Ahlberg PE, Qiao T, Zhu Ya, Zhao W, Jia L. 2016. A Devonian predatory fish provides insights into the early evolution of modern sarcopterygians. *Science advances* 2: e1600154.

Maddison WP. 1991. Squared-Change Parsimony Reconstructions of Ancestral States for Continuous-Valued Characters on a Phylogenetic Tree. *Systematic Biology* 40: 304–315.

Maddison W, Maddison D. 2001. Mesquite: a modular system for evolutionary analysis. *Evolution* 62: 1103-1118.

Marjanović D, Laurin M. 2008. A reevaluation of the evidence supporting an unorthodox hypothesis on the origin of extant amphibians. *Contributions to Zoology* 77: 149–199.

Marjanović D, Laurin M. 2009. The origin(s) of modern amphibians: a commentary. *Evolutionary Biology* 36: 336–338.

Mickevich MF, Johnson MS. 1976. Congruence Between Morphological and Allozyme Data in Evolutionary Inference and Character Evolution. *Systematic Biology* 25: 260–270.

Mickevich MF, Platnick NI. 1989. On the information content of classifications. *Cladistics* 5: 33–47.

Nelson G, Platnick NI. 1991. Three-Taxon Statement: a more precise use of parsimony? *Cladistics* 7: 351–366.

Nelson G, Ladiges PY. 1992. Information content and fractional weight of three-item statements. *Systematic biology* 41: 490–494.

Nelson G, Ladiges PY. 1994. Three-item consensus empirical test of fractional weighting. *Systematics Association Special Volume* 52: 193-193.

Nelson GJ, Ladiges PY. 1996. Paralogy in cladistic biogeography and analysis of paralogy-free subtrees. *American Museum novitates* 3167.

Nelson G. Williams DM, Ebach MC. 2003. A question of conflict: Three-item and standard parsimony compared. *Systematics and Biodiversity* 1: 145–149.

de Pinna MC. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7: 367–394.

de Pinna MC. Bockmann FA, Zaragüeta R. 2016. Unrooted trees discovered independently in philology and phylogenetics: a remarkable case of methodological convergence. *Systematics and Biodiversity* 14: 1-10.

Platnick NI, Humphries CJ, Nelson G, Williams DM 1996. Is Farris Optimization perfect?: three-taxon statements and multiple branching. *Cladistics* 12: 243–252.

Prüfer K,Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. 2010. Computational challenges in the analysis of ancient DNA. *Genome biology* 11: R47.

Pyron RA. 2011. Divergence-time estimation using fossils as terminal taxa and the origins of lissamphibia. *Systematic Biology* 60: 466–481.

Rineau V, Grand A, Zaragüeta R, Laurin M. 2015. Experimental systematics: sensitivity of cladistic methods to polarization and character ordering schemes. *Contributions to Zoology*, 84: 129–148.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53: 131–147.

Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray DL, Rasnitsyn A. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. *Systematic Biology* 61: 973–999.

Sansom RS, Gabbott SE, Purnell MA. 2011. Decay of vertebrate characters in hagfish and lamprey (Cyclostomata) and the implications for the vertebrate fossil record. *Proceedings of the Royal Society B: Biological Sciences* 278: 1150–1157.
Scotland RW. 2000. Taxic homology and three-taxon statement analysis. *Systematic biology* 49: 480–500.
Siebert DJ, Williams DM. 1998. Recycled. *Cladistics* 14: 339–347.
Spencer M, Davidson EA, Barbrook AC, *et al*. 2004. Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology* 227: 503–511.
Swofford D. 2003. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates.
Vinther J, Stein M, Longrich NR, Harper DAT. 2014. A suspension-feeding anomalocarid from the Early Cambrian. *Nature* 507: 496–500.
Willerslev E, Cooper A. 2005. Ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences* 272: 3-16.
Williams DM, Ebach MC. 2005. Drowning by Numbers: Rereading Nelson's "Nullius in Verba". *The Botanical Review* 71: 415-430.
Wright AM, Hillis DM. 2014. Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. *PLOS ONE* 9: e109210.
Zaragüeta R, Bourdon E. 2007. Three-item analysis: Hierarchical representation and treatment of missing and inapplicable data. *Comptes Rendus Palevol* 6: 527–534.
Zaragüeta R, Ung V, Grand A, Vignes-Lebbe R, Cao N, Ducasse J. 2012. LisBeth: New cladistics for phylogenetics and biogeography. *Comptes Rendus Palevol* 11: 563–566.
Zhang C, Stadler T, Klopfstein S, Heath TA, Ronquist F. 2016. Total-evidence dating under the fossilized birth–death process. *Systematic Biology* 65: 228-249.

Received: 25 April 2017
Revised and accepted: 19 December 2017
Published online: 13 April 2018
Editor: R. Sluys

Online supplementary information

S1. Alignment in fasta format of 14 HIV virions and the two (HIV and SIV) outgroups.

S2. Nexus files for PAUP 4.b10 of cladistic analysis of HIV with the SIV outgroup.

S3. Nexus files for PAUP 4.b10 of cladistic analysis of HIV with the SIV outgroup.

S4. 3ta files for Lisbeth 1.0 of cladistic analysis of HIV with the SIV outgroup.

S5. 3ta files for Lisbeth 1.0 of cladistic analysis of HIV with the SIV outgroup.

S6. Alignment of the artificial manuscript dataset comprising the 13 taxa used in the analyses (csv file).

S7. Nexus files for PAUP 4.b10 of the cladistic analysis of manuscripts.

S8. 3ta file for Lisbeth 1.0 of the cladistic analysis of manuscripts.