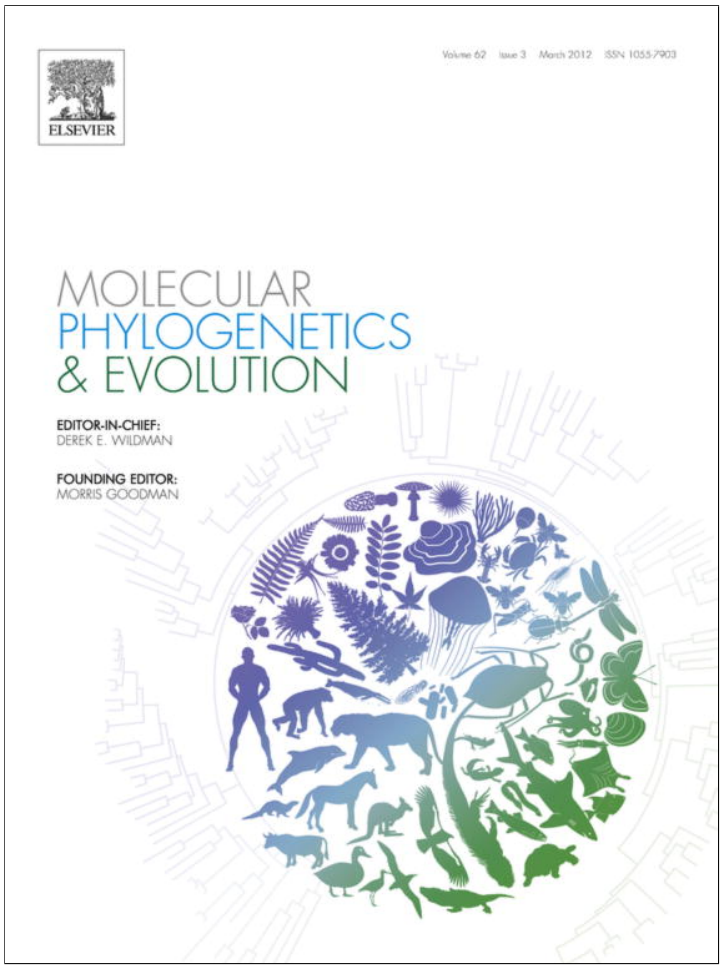

Molecular Phylogenetics and Evolution 62 (2012) 921–927

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

Effect of genetic convergence on phylogenetic inference

Pascal-Antoine Christin [a,b], Guillaume Besnard [c], Erika J. Edwards [b], Nicolas Salamin [a,d,*]

[a] Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland
[b] Department of Ecology and Evolutionary Biology, Brown University, 80 Waterman St., Box G-W, Providence, RI 02912, USA
[c] CNRS-UPS-ENFA, Laboratoire Evolution & Diversité Biologique, UMR 5174, 31062 Toulouse, France
[d] Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 30 July 2011
Revised 30 November 2011
Accepted 2 December 2011
Available online 14 December 2011

Keywords:
Phylogeny
Bias
Evolutionary convergence
Amino acids
Codons
Simulations

ABSTRACT

Phylogenetic reconstructions are a major component of many studies in evolutionary biology, but their accuracy can be reduced under certain conditions. Recent studies showed that the convergent evolution of some phenotypes resulted from recurrent amino acid substitutions in genes belonging to distant lineages. It has been suggested that these convergent substitutions could bias phylogenetic reconstruction toward grouping convergent phenotypes together, but such an effect has never been appropriately tested. We used computer simulations to determine the effect of convergent substitutions on the accuracy of phylogenetic inference. We show that, in some realistic conditions, even a relatively small proportion of convergent codons can strongly bias phylogenetic reconstruction, especially when amino acid sequences are used as characters. The strength of this bias does not depend on the reconstruction method but varies as a function of how much divergence had occurred among the lineages prior to any episodes of convergent substitutions. While the occurrence of this bias is difficult to predict, the risk of spurious groupings is strongly decreased by considering only 3rd codon positions, which are less subject to selection, as long as saturation problems are not present. Therefore, we recommend that, whenever possible, topologies obtained with amino acid sequences and 3rd codon positions be compared to identify potential phylogenetic biases and avoid evolutionarily misleading conclusions.

1. Introduction

Phylogenetic trees are now widely used in many different fields of research from genomics, by studying genes and genome evolution (e.g. Kuzniar et al., 2008; Ravi et al., 2009; Kuo and Ochman, 2010; Mayrose et al., 2010; Nilsson et al., 2010; Slot and Rokas, 2010; Whitney and Garland, 2010), to ecology and conservation, by assessing, for example, evolutionary effects on community assembly (e.g. Mayfield and Levine, 2010; Pillar and Duarte, 2010). A wide range of methods are available to reconstruct phylogenetic trees, and the development of more sophisticated models of molecular evolution has greatly improved the accuracy of phylogenetic inference (Holder and Lewis, 2003). There are, however, many problems that can still make a phylogenetic reconstruction unreliable. A first type of potential source of error is introduced because data sets are, by definition, finite. Random error is therefore likely often present in real data sets (Sanderson et al., 2000; Anderson and Swofford, 2004; Brinkmann et al., 2005; Burleigh and Mathews, 2007; Geuten et al., 2007). It usually leaves detectable signs on phylogenetic reconstructions through a low support for particular nodes and can be reduced by sampling a sufficiently large number of characters (Salamin et al., 2005). Systematic biases, which affect several aspects of the reconstruction process, are more pernicious. The most well-known and studied case is the so-called long-branch attraction problem (e.g. Felsenstein, 1973; Kuhnert and Felsenstein, 1994; Hillis, 1996; Kim, 1998; Sanderson et al., 2000). The systematic error leading to long-branch attraction is due to the over-simplification of the model of evolution used. Although maximum parsimony is particularly prone to this problem, all reconstruction methods are potentially biased. Other systematic errors can arise from any model misspecification (Buckley, 2002).

Thanks to recent developments and the wide use of models to estimate events of positive selection (Yang and Nielsen, 2002; Zhang et al., 2005; Yang, 2006; Anisimova and Kosiol, 2009), the number of genes known to have sites that have evolved under positive selection has drastically increased (e.g. Han et al., 2008; Studer et al., 2008; Singh et al., 2009). Selective processes could therefore represent an additional source of undetected bias in current phylogenetic reconstructions based on protein-coding genes (Edwards, 2009). In particular, it has been shown that similar selection pressures can lead to the repeated emergence of the same phenotype in distant lineages via identical genetic changes (e.g.

* Corresponding author at: Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland. Fax: +41 21 692 4270.
E-mail addresses: pascal-antoine_christin@brown.edu (P.-A. Christin), guillaume.besnard@univ-tlse3.fr (G. Besnard), erika_edwards@brown.edu (E.J. Edwards), nicolas.salamin@unil.ch (N. Salamin).

1055-7903/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2011.12.002

Wood et al., 2005; Zhang, 2006; Castoe et al., 2009; Christin et al., 2010). The recurrence of some amino acid replacements in different lineages has been reported for a wide range of organisms and phenotypes, including the repeated evolution of echolocation in bats and dolphins (Li et al., 2010; Liu et al., 2010), the many origins of C_4 photosynthesis in flowering plants (Christin et al., 2007; Besnard et al., 2009), and convergent evolution of drug and pesticide resistance in microbes and insects (e.g. ffrench-Constant et al., 2004; Liu et al., 2007; Lozovsky et al., 2009). Several of these studies suggested that the convergent adaptive substitutions biased the phylogenetic constructions, tending to group together the genes that had convergent adaptive substitutions, despite their distant relatedness (Stewart et al., 1987; Kriener et al., 2000; Christin et al., 2007; Castoe et al., 2009; Zhang, 2009; Burri et al., 2010; Li et al., 2010). Although this bias is intuitive, its strength and frequency have never been appropriately tested. Moreover, there are still debates on the pertinence of using gene portions, such as 3rd codon positions and non coding regions, that are less subject to selection but usually fast evolving, versus DNA regions that can be affected by selective pressures but are seen as more slowly evolving, such as amino acid sequences (Källersjö et al., 1999; Salamin et al., 2005; Su et al., 2009).

In the present study, we use computer simulations to assess the effect of convergent substitution events on the accuracy of phylogenetic inference. We compare different phylogenetic methods as well as different data partitions, and evaluate the effect of variation in sequence and topology properties on the strength of the phylogenetic biases due to convergent substitutions. Our aims are to test the possibility of a phylogenetic bias under evolutionary convergence at the amino acid level, to identify the conditions in which this bias is likely to occur and to propose guidelines to reduce the effect of genetic convergence on phylogenetic reconstructions.

2. Material and methods

2.1. Tree topologies

The topologies used in this study were inferred from real data sets for the Poaceae (Topologies 1 and 2) and Cyperaceae (Topology 3) to provide realistic tree shapes and depths (Appendix 1). In each topology, evolutionary convergence at the amino acid level was simulated on branches leading to groups with a convergent phenotype (CG), set as genes that were independently co-opted in the C_4 photosynthetic pathway through convergent amino acid replacements in the real data set. Topology 1 is based on two plastid genes (*rbcL* and *ndhF*), one of which underwent convergent amino acid replacements in C_4 species (Christin et al., 2008b). To reduce computational time, a subset of 31 grass species out of the 187 present in the published data set was selected and the tree was reconstructed using MrBayes 3.1 (Ronquist and Huelsenbeck, 2003) under a GTR + Γ + I model. This topology contains four groups of species that independently evolved C_4 photosynthesis, which were defined as CG in the simulations (Appendix 1). Topology 2 is based on nuclear sequences of *ppc-B2* genes that encode phosphoenolpyruvate carboxylase (PEPC; Christin et al., 2007). It contains eight groups of genes that have been independently recruited for a key role in C_4 photosynthesis (Christin et al., 2007; Appendix 1), which were defined as CG in the simulations. The topology of Christin et al. (2007) based on 3rd codon positions and introns was used because it was found congruent with the known species tree (see Christin et al. (2007) for details). Branch lengths for this topology were calculated with PhyML 2.4 (Guindon and Gascuel, 2003) under a GTR + Γ substitution model based on 3rd codon positions only. This was done to avoid strong variations in branch lengths due to different selective pressures in non-C_4 and C_4 genes. Topology 3 is based on 79 PEPC encoding genes (*ppc-1*) of Cyperaceae (Besnard et al., 2009). The topology was inferred from all coding positions, but the branch lengths were also calculated with PhyML 2.4 under a GTR + Γ substitution model based on 3rd positions of codons only. This topology includes five groups of C_4-specific genes that were set as CG in the simulations (Appendix 1).

2.2. Simulation framework

The episodes of convergent evolution at the amino acid level were simulated by adding convergent codons, which were forced to mutate to the exact same amino acid at the base of each CG (hereafter called simply as convergent codons), to a set of regular codons that did not experience such convergent evolution. This allowed us to explore the effect of different proportions of convergent codons on phylogenetic reconstructions, while controlling for all other parameters.

A parametric bootstrap approach was used to simulate codon sequences, which were generated using the evolver software from the PAML package (Yang, 2007). Our simulation framework was inspired from the branch-site model of codon evolution (Yang, 2006) and two categories of codons were generated. Regular codons, exempt of convergent evolution, were simulated under a site model (Nielsen and Yang, 1998) with 95% of codons evolving under purifying selection ($\omega = 0.05$) and 5% evolving under relaxed selection ($\omega = 1$). These values correspond to the parameters estimated for the PEPC data sets (Christin et al., 2007; Besnard et al., 2009). A second set of codons was generated under a branch model (Yang and Nielsen, 2002) using a modified version of evolver (available at http://www.unil.ch/phylo/bioinformatics/other-software.html). These codons evolved under purifying selection ($\omega = 0.05$) in most branches of the phylogenetic tree but changed to codons encoding the same amino acid in branches at the base of each monophyletic CG. The amino acid selected as convergent at each site and the synonymous codon used for each monophyletic CG were drawn randomly according to the amino acid and codon frequencies of the original data, respectively. This represents an episode of extreme substitution convergence that is associated with the evolution of a similar phenotype involving changes in the properties of the encoded protein (Christin et al., 2010). After each convergence event, codons were allowed to diverge again along the descendant branches under purifying selection ($\omega = 0.05$). The total number of codons used for phylogenetic analyses was assembled by mixing a variable proportion of convergent codons with regular codons. The simulated data sets always contained 400 amino acids (i.e. 1200 nucleotides), but the proportion of convergent codons included in the datas ets varied from 0 (i.e. all codons were non-convergent) to 10% (i.e. 40 out of the 400 codons were convergent) by steps of 0.25%. For each topology, 100 replicates were performed for each proportion.

For illustration purpose, one simulated data set was investigated in more details. A simulation including 32 convergent codons, a proportion that strongly biased the phylogenetic reconstructions (see below), generated with Topology 2, was randomly selected to explore how phylogenetic information from sites is distributed along the simulated sequence. Using PAUP* 4.0b10 (Swofford, 2002), the log-likelihood at each site was calculated under a GTR + Γ model of substitution, for both the inferred topology (obtained with PhyML under a HKY model) that was biased toward grouping all CG and the topology used to generate the data. The difference between the log-likelihoods was calculated by subtracting the log-likelihood of the true topology to the log-likelihood of the inferred topology. Negative values indicate that the site favors the inferred topology.

2.3. Comparisons of different phylogenetic reconstruction methods

Several phylogenetic inference methods were used on each simulation replicate to assess the effect of tree reconstruction

algorithms. However, due to the computational burden involved, it was not possible to assess all possible heuristics and several representative programs were selected for this task.

For nucleotide sequences, neighbor-joining trees were obtained using dnadist and neighbor (Felsenstein, 2005), under a F84 substitution model. Maximum parsimony trees were inferred using dnapars (Felsenstein, 2005) with heuristic rearrangements made on one of the best trees only. The consensus of all maximum parsimony trees was then considered. Under a maximum-likelihood criterion, trees were inferred using dnaml (phylip 3.6; Felsenstein, 2005) under a HKY model, and using PhyML 2.4 under either a HKY or a GTR + Γ + I substitution model. For each simulation replicate, all these reconstruction methods were used to infer trees from all codon positions as well as from 3rd codon positions only. The nucleotide sequences were also translated into amino acids. In this case, neighbor-joining trees were obtained using protdist and neighbor (phylip 3.6; Felsenstein, 2005) under a JTT substitution model. Maximum parsimony trees were inferred using protpars (Felsenstein, 2005). Finally, maximum-likelihood trees were inferred using PhyML under both a JTT and a JTT + Γ substitution models.

The number of CG on the inferred phylogenetic tree was compared with the number of CG in the topology used to generate the sequences. A reduction of the number of CG in the topologies inferred from simulated data indicates a bias toward grouping CG, despite their different origins. We also measured the topological distance between each phylogenetic tree inferred from simulated data sets and the true topology used to generate the sequences. The method of Penny and Hendy (1985) was computed for this purpose as implemented in the APE package of R (Paradis et al., 2004). This distance is measured as twice the number of internal branches defining different bipartitions of the tips (Penny and Hendy, 1985).

2.4. Parameters affecting the phylogenetic bias

Further simulations were performed only on Topology 1 and under maximum-likelihood using PhyML 2.4 with a HKY substitution model to explore possible interactions between the proportion of convergent codons and matrix size or tree depth. The sequence length was first modified by varying the total number of codons from 125 to 6000 while maintaining the original tree depth. Secondly, different tree depths were investigated by multiplying each branch of the phylogeny by a factor ranging from 0.25 to 13.5, increasing in 0.25 increments. For these simulations, sequence length was set to 1000 codons, to increase the amount of information. For each data set, we increased the number of convergent codons until 80 of the 100 simulation replicates reconstructed a single CG instead of the four present in the true Topology 1 (Appendix 1).

A similar procedure was used to investigate how the phylogenetic distance between two CG influences the bias due to convergent codons. All possible pairs of internal branches of Topology 1 were successively defined as CG unless they corresponded to sister-groups or were nested within one another. For each pair, the minimal number of convergent codons to lead to the inference of one CG group in 80% of the simulations was calculated as described above. Sequence length was set to 1000 codons and the original tree depth was kept. The distance separating each pair of CG was determined as the sum of branch lengths to join the stem group nodes of the two CG. Additionally, the same procedure was repeated with (i) a modified topology based on Topology 1 (Topology 4, Appendix 1), which contains only a subset of the tips present but with identical branch lengths, and (ii) Topology 2 with branches divided by 10 to obtain expected number of substitutions per codon that were comparable with those of Topology 1. To reduce the computation time, only 28 nodes of Topology 2, drawn randomly from all internal nodes, were successfully set as CG in combinations of two.

2.5. Relevance of character partitions with varying tree depths

To assess the relevance of the different data partitions (all nucleotide positions, 3rd codon positions and amino acids) with different tree depths, we inferred phylogenetic trees under maximum likelihood in PhyML 2.4 with a HKY substitution model for nucleotides and JTT for amino acids. Data sets were simulated without convergent codons but by multiplying all branches of the phylogeny by a factor ranging from 1 to 100. For each simulated data set, the topological distance between the inferred phylogeny and the true topology was computed using the method of Penny and Hendy (1985). This procedure was repeated with Topologies 1, 2 and 3.

3. Results and discussion

3.1. Phylogenetic bias resulting from genetic convergence

The simulations showed that a small number of convergent codons can strongly bias the phylogenetic inferences. With Topology 1, for instance, a single convergent codon out of 400 was sufficient to bias the phylogenetic inference on amino acids (Fig. 1, Appendix 2). The strength of the bias varied greatly among topologies and as a function of the number of convergent codons (Fig. 1; Appendices 2, 3 and 4). The number of convergent codons required to bias 80% of the simulations towards inferring a single CG was 10 (2.5%), 25 (6.3%) and 37 (9.3%) using maximum likelihood (HKY model) on all codon positions for Topologies 1, 2 and 3, respectively (Table 1). The non-synonymous positions in convergent codons are identical in the different CG following the episode of convergence. They continue to carry a very strong misleading phylogenetic signal despite the independent substitutions occurring on these nucleotides along the branch lengths following the convergent event. On the other hand, the phylogenetic signal carried by the regular codons is spread among the multiple branches of the topology. While these regular codons give an accurate phylogenetic signal, only a very small proportion of them will convey information to correctly place the different CG within the phylogenetic tree (Fig. 2). Thus, the misleading information of the convergent codons will quickly exceed the information of regular codons as their proportion increases (Table 2; Fig. 2). The different reconstruction methods led to similar patterns, although more thorough heuristic searches (as with dnaml) tended to reduce the bias slightly (Table 1; Appendices 2, 3 and 4). The distance between the true and inferred topologies increased with inclusion of more convergent codons, mirroring the inferred number of CG (Fig. 1, Appendices 2, 3 and 4).

Overall, the bias was particularly strong whenever amino acid sequences were used to infer the phylogenetic tree, moderate with all three codon positions and very low when only the 3rd codon positions were considered (Fig. 1). For inferences based on amino acid sequences, the number of convergent codons required to lead to 80% of the simulations inferring a single CG was 4 (1.0%), 10 (2.5%) and 13 (3.3%) using maximum likelihood (JTT model) for Topologies 1, 2 and 3, respectively (Table 1). Here again, the different inference methods yielded similar results, which shows that the bias will be highly prevalent when using such type of data. The data sets that allowed the best inference despite the presence of convergent codons was the 3rd codon positions, which required 35 (8.8%), 91 (22.8%) and 117 (29.3%) convergent codons to lead maximum likelihood (HKY model) to infer a single CG on Topologies 1, 2 and 3, respectively. When using amino acid sequences, the phylogenetic signal of regular codons is reduced because most mutations in codons under purifying selection occur on synonymous sites. As a consequence, trees inferred from amino acid sequences are more subject to phylogenetic bias in presence of

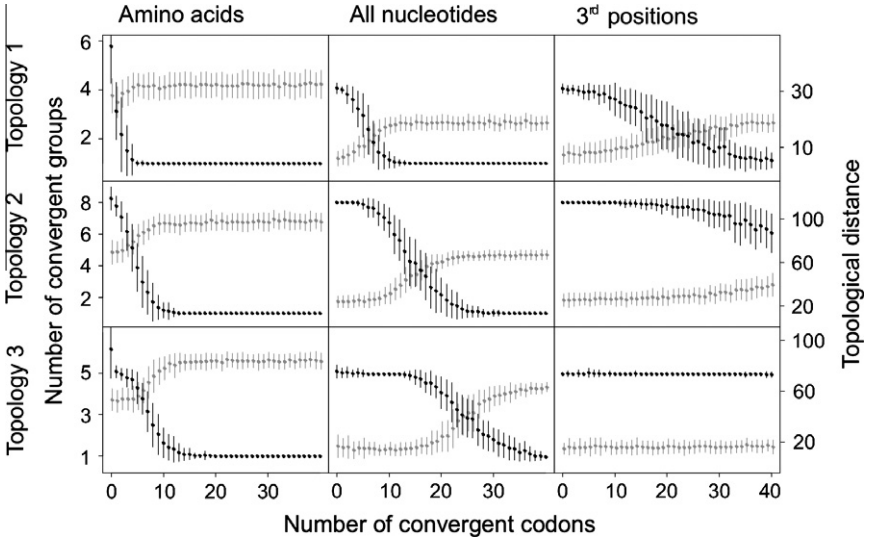

Fig. 1. Phylogenetic inference as a function of the number of convergent codons. The number of CG inferred from the simulated data sets (black points, scale on the left) and the topological distance between the inferred topology and the true topology (gray points, scale on the right) are indicated for different numbers of convergent codons, for the three topologies and for inferences based on amino acid sequences, all nucleotide positions and 3rd codon positions. Each point is the mean over 100 replicates and the standard deviation is indicated. The HKY substitution model as implemented in PhyML was used for all nucleotides and 3rd codon positions analyses and the JJT model in PhyML was used for amino acid sequences.

Table 1
Minimal number of convergent codons leading to only one CG in more than 80% of the simulations. One hundred replicates were performed for all phylogenetic reconstruction methods and data partitions, for the three different topologies. Percent of the codon sequences corresponding to the number of convergent codons is indicated in parentheses. NJ = neighbor-joining, MP = maximum-parsimony, ML = maximum-likelihood.

Partition	Phylogenetic method	Topology 1	Topology 2	Topology 3
Amino acids	NJ	3 (0.8)	8 (2.0)	11 (2.8)
	MP	4 (1.0)	11 (2.8)	13 (3.3)
	ML, JTT	4 (1.0)	10 (2.5)	13 (3.3)
	ML, JTT + G	4 (1.0)	10 (2.5)	13 (3.3)
All nucleotides	NJ	11 (2.8)	28 (7.0)	43 (10.8)
	MP	10 (2.5)	24 (6.0)	34 (8.5)
	ML, dnaml	11 (2.8)	27 (6.8)	48 (12.0)
	ML, HKY	10 (2.5)	25 (6.3)	37 (9.3)
	ML, GTR + G + *I*	10 (2.5)	23 (5.8)	36 (9.0)
3rd codon positions	NJ	44 (11.0)	106 (26.5)	146 (36.5)
	MP	36 (9.0)	85 (21.3)	112 (28.0)
	ML, dnaml	43 (10.8)	109 (27.3)	157 (39.3)
	ML, HKY	35 (8.8)	91 (22.8)	117 (29.3)
	ML, GTR + G + *I*	36 (9.0)	94 (23.5)	119 (29.8)

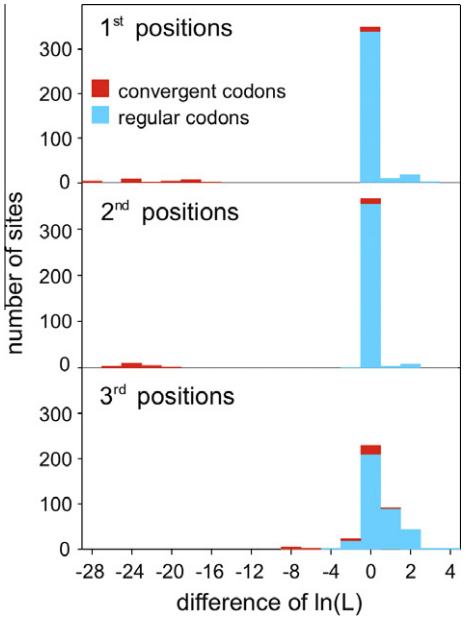

Fig. 2. Repartition of the phylogenetic signal for one data set biased by the presence of convergent codons. The data set was generated with Topology 1 and includes 32 convergent codons out of 400. For each nucleotide site, the difference in log-likelihood between the inferred (biased) and the true topologies was calculated. A negative value indicates that the site favors the biased topology. The distribution of these differences is presented for 1st, 2nd and 3rd codon positions, separately for regular codons (in blue) and convergent codons (in red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

genetic convergence. In contrast, phylogenetic signal of regular codons is maximal on 3rd codon positions (Table 2; Fig. 2), which contain mostly synonymous sites and are the most variable sites under purifying selection. At the same time, the misleading effect of convergent codons is very low on these positions because most of these sites are not involved in the convergence at the amino acid level (Table 2; Fig. 2). Consequently, the phylogenetic bias is very strongly decreased when considering 3rd codon positions only.

3.2. Parameters affecting the phylogenetic bias

The number of convergent codons necessary to mislead the analyses increased linearly with the total number of codons in the sequence (Fig. 3). This means that the proportion of convergent codons that biases the phylogenetic analyses is constant for a given tree topology and selection regime. This is probably because both the accurate phylogenetic signal of regular codons and the misleading signal of convergent codons increase linearly with the length of the sequence. Consequently, considering longer

Table 2
Repartition of the phylogenetic signal in one data set biased by the presence of convergent codons. This data set was generated with Topology 1 and includes 32 convergent codons out of 400. The sum of the differences in log-likelihood between the inferred and the true topologies is indicated for the different positions of both regular and convergent codons. Negative values indicate that the partition favors the topology that groups all CG.

	1st positions	2nd positions	3rd positions	Total
Regular codons	91.79	36.91	331.06	459.75
Convergent codons	−486.1	−488.98	−51.54	−1026.72

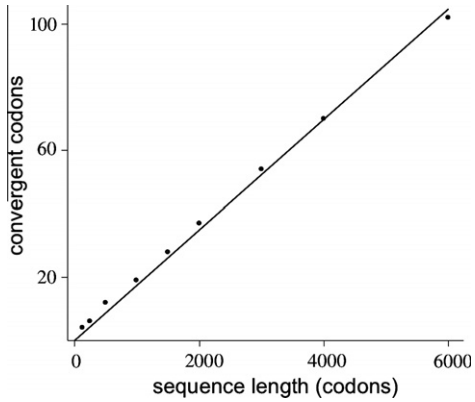

Fig. 3. Number of convergent codons that leads to a strong bias for different total number of codons. For different sequence lengths (total number of codons), the minimal number of convergent codons to obtain only one CG in more than 80% of replicates was determined. The relationship is linear (r-squared = 0.998), with a slope of 0.0175 (black line).

sequences is unlikely to decrease the bias if the convergent codons are spread homogeneously along the DNA region studied.

The bias due to convergent codons decreased with the tree depth (Fig. 4). The correct phylogenetic signal brought by regular codons is more important on trees with higher expected rates of substitution, while the amount of misleading signal due to convergent codons remains unchanged (the number of convergent substitutions is unchanged). Consequently, more convergent codons are required to bias the phylogenetic inference. The plateau might be associated with a stagnation of the number of phylogenetically informative substitutions when the sequences become saturated by an excess of mutations.

3.3. Distance between the convergent groups is a key parameter

Our simulations indicated that the number of convergent codons needed to bias the phylogenetic inferences is a function of the distance separating the stem nodes of the two CG (Fig. 5). The relationship is not linear and resembles the relationship seen for the total tree depth (Fig. 4). The plateau was however not reached with the considered range of distances between CG. The same relationship was also observed with Topology 4, indicating that the ratio between the number of species and the sum of branch lengths over the tree was not affecting the phylogenetic bias. In addition, pairs of CG belonging to a completely different topology (Topology 2) behaved in the exact same way (Fig. 5). This suggests that the distance separating two branches where convergent substitutions occur is the critical determinant of the strength of the phylogenetic bias. Other parameters could probably also affect the phylogenetic bias, but their effect seems small on the range of distances considered. When the distance between CG is larger, the regular codons have accumulated a greater amount of phylogenetic signal that helps to place the CG correctly, and a higher number of convergent codons is needed to break this relationship and lead to a biased inference.

The distance between the convergent groups is a likely explanation for the differences in the strength of the bias due to convergent codons between the phylogenetic trees for which convergent substitutions were reported. Indeed, the grass PEPC phylogenetic tree (Topology 2) was strongly affected by the presence of about 20 sites that underwent convergent substitutions (Christin et al., 2007). On the other hand, the sedge PEPC phylogeny (Topology 3) was not biased by a similar number of convergent substitutions (Besnard et al., 2009). These differences are also represented in the output of our simulations, where the bias is more important for Topology 2 (Table 1; Fig. 1). The distance separating the stems of the CG was smaller in Topology 2 than in Topology 3 (maximal distance was 0.39 substitutions per codon versus 0.52). This difference is likely due to contrasted divergence times prior to the convergence events. Indeed, while the main C_4 clades of Poaceae diverged around 20 million years before the episode of convergence (Christin et al., 2008a), the C_4 clades of Cyperaceae diverged up to 40 million years between the convergence episodes (Besnard et al., 2009). Thus, more phylogenetically informative mutations have been accumulated before the convergent substitutions in Cyperaceae,

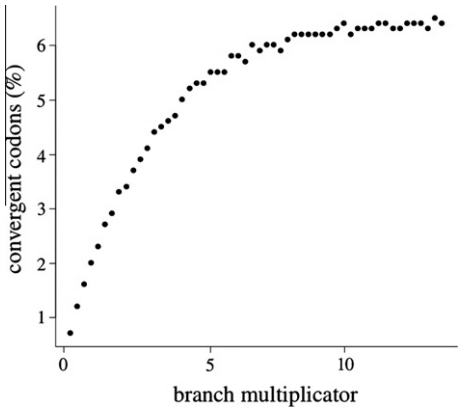

Fig. 4. Number of convergent codons that leads to a strong bias as a function of the tree depth. Different tree depths were obtained by multiplying the length of each branch by a varying factor (branch multiplicator). For each value of this factor, the minimal proportion of convergent codons (in percent) to obtain only one CG in more than 80% of replicates was determined.

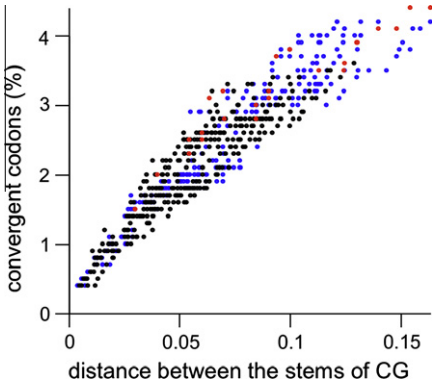

Fig. 5. Number of convergent codons that leads to a strong bias for all possible pairs of convergent groups. For each possible pair of CG, the minimal proportion (in percent) of convergent codons for 80% of simulations leading to only one CG was computed, and is plotted against the sum of branches separating the stem group nodes of the two CG. Points in blue, black and red are for Topologies 1, 2 and 4, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

making Topology 3 more robust to convergent codons. Overall, the importance of distance in determining the strength of the phylogenetic bias means that phylogenies are more likely to be biased when groups with convergent phenotypes are more closely related.

3.4. Relevance of different partitions for various tree depths

The topologies used in these simulations represent plastid genes that diverged during approximately 60 million years (Topology 1; Christin et al., 2008a) and nuclear genes that diverged during roughly the same period (Topologies 2 and 3; Christin et al., 2007; Besnard et al., 2009). The time scale investigated is therefore largely shorter than studies concerned with reconstructions of large taxonomic scale phylogenetic trees. This raises questions about the pertinence of the different partitions for deep phylogenetic trees. Increasing the expected number of substitutions per branch by multiplying all branches of the phylogenetic tree by a factor gradually decreases the accuracy of 3rd codon positions, and leads to a high rate of errors when branches are more than 3–10 times those of the initial phylogeny (Appendix 5). On the other hand, the accuracy of amino acid sequences, which is largely below nucleotides for the initial phylogenetic trees, increases with the multiplication of all branch lengths, making this partition useful on large taxonomic scales (Appendix 5). The implication of these patterns is that 3rd codon positions are pertinent only for studies investigating genes that diverged in recent geological times. On the other hand, reconstructions of deep phylogenies are unlikely to be biased by the presence of convergent codons since an important divergence prior to the convergence event strongly reduces the risk of spurious grouping (Fig. 5). It should be noted that we assume here that episodes of convergent selective pressure is happening once over a short time interval. If evolutionary processes leading to the convergent pattern are recurrent or affect lineages over longer periods of times, the presence of the bias might be stronger and thus still affect deep phylogenetic reconstructions based on amino acids. We have however not tested this case in our simulations.

4. Conclusions

Our simulations have shown that even a small proportion of convergent substitutions can strongly affect the accuracy of phylogenetic reconstructions of genes that diverged in the last hundred million years. The proportion of convergent codons that lead to an important bias varies among phylogenetic trees, mainly as a function of the number of mutations accumulated before the convergent substitutions. Compared to phylogenetic trees inferred from nucleotide sequences, the bias will be stronger when inferences are based on amino acid sequences. On the other hand, considering only 3rd codon positions drastically reduces the phylogenetic bias. Since reports of convergent substitutions are accumulating (Christin et al., 2010), the phylogenetic bias shown in this study could concern many different genes in a wide range of organisms. Currently, no *a priori* diagnostic tool is available to identify such genes and we strongly suggest comparing the phylogenetic signal given by amino acid sequences and 3rd codon positions in phylogenetic studies concerned with the phylogenetic inference of single gene trees diverged relatively recently. An important difference in the inferred topologies could be due to convergent substitutions and the data set should be carefully reevaluated to avoid incorrect conclusions due to misleading phylogenetic trees.

Acknowledgments

Support for computational resources was provided by the Vital-IT facilities from the Swiss Institute of Bioinformatics. This work was supported by the Marie Curie IOF 252568 fellowship to PAC and the Swiss National Science Foundation grant 3100AO-116412 to NS.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2011.12.002.

References

Anderson, F.E., Swofford, D., 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Mol. Phylogenet. Evol. 33, 440–451.
Anisimova, M., Kosiol, C., 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol. Biol. Evol. 26, 255–271.
Besnard, G., Muasya, A.M., Russier, F., Roalson, E.H., Salamin, N., Christin, P.-A., 2009. Phylogenomics of C_4 photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence. Mol. Biol. Evol. 26, 1909–1919.
Brinkmann, H., van der Giezen, M., Zhou, Y., de Raucourt, G.P., Philippe, H., 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54, 743–757.
Buckley, T.R., 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst. Biol. 51, 509–523.
Burleigh, J.G., Mathews, S., 2007. Assessing systematic error in the inference of seed plant phylogeny. Int. J. Plant Sci. 168, 125–135.
Burri, R., Salamin, N., Studer, R.A., Roulin, A., Fumagalli, L., 2010. Adaptive divergence of ancient gene duplicates in the avian MHC class II beta. Mol. Biol. Evol. 27, 2360–2374.
Castoe, T.A., de Koning, A.P.J., Kim, H.M., Gu, W., Noonan, B.P., Naylor, G., Jiang, Z.J., Parkinson, C.L., Pollock, D.D., 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. USA 106, 8986–8991.
Christin, P.-A., Salamin, N., Savolainen, V., Duvall, M.R., Besnard, G., 2007. C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes. Curr. Biol. 17, 1241–1247.
Christin, P.-A., Besnard, G., Samaritani, E., Duvall, M.R., Hodkinson, T.R., Savolainen, V., Salamin, N., 2008a. Oligocene CO_2 decline promoted C_4 photosynthesis in grasses. Curr. Biol. 18, 37–43.
Christin, P.-A., Salamin, N., Muasya, A.M., Roalson, E.H., Russier, F., Besnard, G., 2008b. Evolutionary switch and genetic convergence on *rbcL* following the evolution of C4 photosynthesis. Mol. Biol. Evol. 25, 2361–2368.
Christin, P.-A., Weinreich, D.M., Besnard, G., 2010. Causes and evolutionary significance of genetic convergence. Trends Genet. 26, 400–405.
Edwards, S.V., 2009. Natural selection and phylogenetic analysis. Proc. Natl. Acad. Sci. USA 106, 8799–8800.
Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods forestimating evolutionary trees from data on discrete characters. Syst. Zool. 22, 240–249.
Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) Version 3.6. Distributed by The Author. Department of Genome Sciences, University of Washington, Seattle.
ffrench-Constant, R.H., Daborn, P.J., Le Goff, G., 2004. The genetics and genomics of insecticide resistance. Trends Genet. 20, 163–170.
Geuten, K., Massingham, T., Darius, P., Smets, E., Goldman, N., 2007. Experimental design criteria in phylogenetics: where to add taxa. Syst. Biol. 56, 609–622.
Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.
Han, M.W., Demuth, J.P., McGrath, C.L., Casola, C., Hahn, M.W., 2008. Adaptive evolution of young gene duplicates in mammals. Genome Res. 19, 859–867.
Hillis, D.M., 1996. Infering complex phylogenies. Nature 383, 130–131.
Holder, M., Lewis, P.O., 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. 4, 275–284.
Källersjö, M., Albert, V.A., Farris, J.S., 1999. Homoplasy increases phylogenetic structure. Cladistics 15, 91–93.
Kim, J., 1998. Large-scale phylogenies and measuring the performance of phylogenetic estimators. Syst. Biol. 47, 43–60.
Kriener, K., O'hUigin, C., Tichy, H., Klein, J., 2000. Convergent evolution of major histocompatibility complex molecules in humans and New World monkeys. Immunogenetics 51, 169–178.
Kuhnert, M.K., Felsenstein, J., 1994. Simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.
Kuo, C.H., Ochman, H., 2010. The extinction dynamics of bacterial pseudogenes. PLoS Genet. 6, e1001050.
Kuzniar, A., van Ham, R.C., Pongor, S., Leunissen, J.A., 2008. The quest for orthologs: finding the corresponding gene across genomes. Trends Genet. 24, 539–551.
Li, Y., Liu, Z., Shi, P., Zhang, J., 2010. The hearing gene Prestin unites echolocating bats and whales. Curr. Biol. 20, R55–R56.
Liu, W.J., Harrison, D.K., Chalupska, D., Gornicki, P., O'Donnell, C.C., Adkins, S.W., Haselkorn, R., Williams, R.R., 2007. Single-site mutations in the carboxyltransferase domain of plastid acetyl-CoA carboxylase confer resistance to grass-specific herbicides. Proc. Natl. Acad. Sci. USA 104, 3627–3632.
Liu, Y., Rossiter, S.J., Han, X.Q., Cotton, J.A., Zhang, S.Y., 2010. Cetaceans on a molecular fast track to ultrasonic hearing. Curr. Biol. 20, 1834–1839.

Lozovsky, E.R., Chookajorn, T., Brown, K.M., Imwong, M., Shaw, P.J., Kamchonwongpaisan, S., Neafsey, D.E., Weinreich, D.M., Hartl, D.L., 2009. Stepwise acquisition of pyrimethamine resistance in the malaria parasite. Proc. Natl. Acad. Sci. USA 106, 12025–12030.

Mayfield, M.M., Levine, J.M., 2010. Opposing effects of competitive exclusion on the phylogenetic structure of communities. Ecol. Lett. 13, 1085–1093.

Mayrose, I., Barker, M.S., Otto, S.P., 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. Syst. Biol. 59, 132–144.

Nielsen, R., Yang, Z.H., 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148, 929–936.

Nilsson, M.A., Churakov, G., Sommer, M., Van Tran, N., Zemann, A., Brosius, J., Schmitz, J., 2010. Tracking marsupial evolution using archaic genomic retroposon insertions. PLoS Biol. 8, e1000436.

Paradis, A., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Penny, D., Hendy, M.D., 1985. The use of tree comparison metrics. Syst. Zool. 34, 75–82.

Pillar, V.D., Duarte, L.D.S., 2010. A framework for metacommunity analysis of phylogenetic structure. Ecol. Lett. 13, 587–596.

Ravi, V., Lam, K., Tay, B.H., Tay, A., Brenner, S., Venkatesh, B., 2009. Elephant shark (*Callorhinchus milii*) provides insights into the evolution of Hox gene clusters in gnathostomes. Proc. Natl. Acad. Sci. USA 106, 16327–16332.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

Salamin, N., Hodkinson, T.R., Savolainen, V., 2005. Towards building the tree of life: a simulation study for all angiosperm genera. Syst. Biol. 54, 183–196.

Sanderson, M.J., Wojciechowski, M.F., Hu, J.M., Khan, T.S., Brady, S.G., 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. 17, 782–797.

Singh, N.D., Larracuente, A.M., Sackton, T.B., Clark, A.G., 2009. Comparative genomics on the Drosophila phylogenetic tree. Annu. Rev. Ecol. Evol. Syst. 40, 459–480.

Slot, J.C., Rokas, A., 2010. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. Proc. Natl. Acad. Sci. USA 107, 10136–10141.

Stewart, C.B., Schilling, J.W., Wilson, A.C., 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature 330, 401–404.

Studer, R.A., Penel, S., Duret, L., Robinson-Rechavi, M., 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. Genome Res. 18, 1393–1402.

Su, Z., Xu, L., Gu, Z., 2009. Origins of digestive RNases in leaf monkeys are an open question. Mol. Phylogenet. Evol. 53, 610–611.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Whitney, K.D., Garland Jr., T., 2010. Did genetic drift drive increases in genome complexity? PLoS Genet. 6, e1001080.

Wood, T.E., Burke, J.M., Rieseberg, L.H., 2005. Parallel genotypic adaptation: when evolution repeats itself. Genetica 123, 157–170.

Yang, Z.H., 2006. Computational Molecular Evolution. Oxford University Press, Oxford, England.

Yang, Z.H., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Yang, Z.H., Nielsen, R., 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19, 908–917.

Zhang, J.Z., 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. Nat. Genet. 38, 819–823.

Zhang, J.Z., Nielsen, R., Yang, Z.H., 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22, 2472–2479.

Zhang, J.Z., 2009. Phylogenetic evidence for parallel adaptive origins of digestive RNases in Asian and African leaf monkeys: a response to Xu et al. (2009). Phylogenet. Evol. 53, 608–609.