Bioinformatics, 35(11), 2019, 1862–1869
doi: 10.1093/bioinformatics/bty800
Advance Access Publication Date: 25 October 2018
Original Paper

OXFORD

Phylogenetics

Measuring phylogenetic signal between categorical traits and phylogenies

Rui Borges[1,2,3], João Paulo Machado[1], Cidália Gomes[1], Ana Paula Rocha [2,3] and Agostinho Antunes [1,2,*]

[1]CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, Terminal de Cruzeiros do Porto de Leixões, 4450-208 Matosinhos, Portugal, [2]Department of Biology, Faculty of Sciences of the University of Porto, FCUP, 4169-007 Porto, Portugal and [3]CMUP, Centre of Mathematics of the University of Porto, 4169-007 Porto, Portugal

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on October 19, 2017; revised on August 18, 2018; editorial decision on September 9, 2018; accepted on October 24, 2018

Abstract

Motivation: Determining whether a trait and phylogeny share some degree of phylogenetic signal is a flagship goal in evolutionary biology. Signatures of phylogenetic signal can assist the resolution of a broad range of evolutionary questions regarding the tempo and mode of phenotypic evolution. However, despite the considerable number of strategies to measure it, few and limited approaches exist for categorical traits. Here, we used the concept of Shannon entropy and propose the δ statistic for evaluating the degree of phylogenetic signal between a phylogeny and categorical traits.

Results: We validated δ as a measure of phylogenetic signal: the higher the δ-value the higher the degree of phylogenetic signal between a given tree and a trait. Based on simulated data we proposed a threshold-based classification test to pinpoint cases of phylogenetic signal. The assessment of the test's specificity and sensitivity suggested that the δ approach should only be applied to 20 or more species. We have further tested the performance of δ in scenarios of branch length and topology uncertainty, unbiased and biased trait evolution and trait saturation. Our results showed that δ may be applied in a wide range of phylogenetic contexts. Finally, we investigated our method in 14 360 mammalian gene trees and found that olfactory receptor genes are significantly associated with the mammalian activity patterns, a result that is congruent with expectations and experiments from the literature. Our application shows that δ can successfully detect molecular signatures of phenotypic evolution. We conclude that δ represents a useful measure of phylogenetic signal since many phenotypes can only be measured in categories.

Availability and implementation: https://github.com/mrborges23/delta_statistic.

Contact: aantunes@ciimar.up.pt

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Phylogenetic signal represents the tendency for closely related species to resemble each other more than less related taxa, as the result of shared evolutionary history. The concept of phylogenetic signal has been used to answer a wide range of questions about how, when and why different traits evolve: e.g. solving primate behavior, ecology and life history (Kamilar and Cooper, 2013), assessing pathogen-host interactions (Antunes *et al.*, 2008; Gilbert and Webb, 2007), predicting patterns of ecological similarity (Losos, 2008), understanding niche dynamics (Pearman *et al.*, 2008), measuring extinction risk in mammals (Fritz and Purvis, 2010), evaluating species vulnerability to climate change (Thuiller *et al.*, 2011), reconstructing ancient language history (Dunn, 2005), improving phylogenetic analyses (Simmons and Ochoterena, 2000) and solving phylogenetic

© The Author(s) 2018. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022

conflicts (Burleigh and Mathews, 2004). Because of the central importance of phylogenetic signal in comparative phylogenetics, several methods to measure it have been proposed over the last 20 years.

The first attempt to estimate phylogenetic signal was proposed by Pagel in 1999 (Pagel, 1999). Pagel's λ has been extensively used in phylogenetic studies and it is defined under the Brownian motion model of evolution (i.e. a Gaussian random process in which the trait variance accumulates proportionally to evolutionary time) (Pagel, 1999). Pagel's λ ranges between zero and one: zero indicating independent evolution between the trait and the phylogeny, and one indicating co-evolution according to the expectation of the Brownian motion (Pagel, 1999). Pagel's λ expresses the transformation of the internal branch lengths that best predicts the distribution of traits: when zero, the internal structure of the tree is entirely eliminated (yielding a star phylogeny in which all the species evolved independently); in contrast, when λ approaches 1 the internal branch lengths stay almost unchanged, suggesting the trait evolved according to the phylogeny.

Several other statistics of phylogenetic signal followed Pagel's λ, each differing in terms of the approach to measure phylogenetic signal (autocorrelation or evolutionary, model-based or not), the inferential framework and the type of data they can cope with (Table 1). As these statistics follow alternative approaches, they all measure different aspects of phylogenetic signal and, consequently, respond differently to inaccurate phylogenetic information, low sample sizes and the absence of branch length information (Münkemüller *et al.*, 2012).

Most diagnostics of phylogenetic signal were created to analyze continuous data, while few are able to deal with categoric data: the *D* statistic and that presented in this work. The main reason for this is the difficulty to translate the Pagel's principle to categorical data: it is impossible to calculate variances and co-variances with categorical traits. For example, the *D* statistic avoids working directly with categorical traits by defining a trait discretization based on a continuous trait that evolves under the Brownian motion (Fritz and Purvis, 2010). Therefore, the *D* statistic cannot be used in studies where categorical traits do not evolve according to a Brownian motion threshold model (Felsenstein, 2005). Although few statistics of phlylogenetic signal exist for categorical data, these types of variables are common in evolutionary studies, as several aspects of the species' ecology, physiology, morphology and behavior can only be characterized in categories.

In this work, we propose a phylogenetic analog of the Shannon entropy for measuring the degree of phylogenetic signal between a categorical trait and a phylogeny, δ. We used simulated data to test the performance of our statistic, which validates the usefulness of δ to test for phylogenetic signal. An application of δ to 14 360 mammalian gene trees shows its practicality to identify biological processes underlying the evolution of complex phenotypes.

2 Materials and methods

2.1 Introducing and calculating the δ statistic

Translating the principle of phylogenetic signal to categorical data is challenging because one cannot calculate standard summary statistics such as expected values, variances and co-variances. We circumvent this difficulty by, instead, measuring the entropy contained in ancestral inferences. Ancestral reconstructions using categorical data return the probability of each trait category occurring in each node. It is our expectation that the better a phylogeny is associated with a given trait, the better it is able to retrace the trait's evolution, or in probabilistic terms, to infer the ancestral states with minimal uncertainty. To assess the uncertainty of the ancestral inferences, we used the concept of Shannon's entropy from Information Theory (Shannon, 1948), which measures the expected information contained in probabilistic messages. Figure 1 schematically represents the kind of relationship between ancestral inferences, entropy and phylogenetic signal, which will be introduced and discussed in the following sections of the manuscript.

Consider a k-state categorical trait for which ancestral reconstructions have been performed for the N nodes of a given phylogeny with s species. The probability of observing the state i in the node j is defined as p_i^j (state probabilities). Our method uses previously calculated node probabilities that can be obtained by any method returning probability vectors for the ancestral reconstructions: e.g. time-continuous discrete-trait Markov chain models with maximum likelihood or Bayesian inferential frameworks (but not parsimony). Our strategy to calculate the quantity of information in the node probabilities included transforming p_i^j using a linear version of the Shannon entropy (Equation 1)

$$e_i^j = \begin{cases} p_i^j & \text{if } p_i^j \leq 1/k \\ \dfrac{1}{1-k}p_i^j - \dfrac{1}{1-k} & \text{if } p_i^j > 1/k \end{cases}. \quad (1)$$

e_i^j is the ith state entropy and measures the quantity of information given by the probability of state i to occur in node j (Equation 1): when p_i^j is either 0 or 1 the entropy becomes 0, while the maximum entropy is obtained when $p_i^j = 1/k$. The last assumption becomes clear by noting that the scenario of absolute uncertainty occurs when all the k-states are inferred with equal probability (i.e. a uniform vector: $p_i^j = 1/k$ for all i).

The jth node entropy e^j can be obtained summing up the state entropies (Equation 2)

$$e^j = \sum_{i=1}^{k} e_i^j. \quad (2)$$

e^j varies between 0 and 1, which corresponds to situations of absolute certainty (one of the p_i^j is 1 and all the others are 0) and

Table 1. Statistics of phylogenetic signal

Statistics	Approach	Model based	Statistical framework	Data	References
Moran's *I*	Autocorrelation	No	Permutation	Continuous	Moran (1950)
Abouheif's *C*	Autocorrelation	No	Permutation	Continuous	Abouheif (1999)
Pagel's λ	Evolutionary	Yes	Maximum Likelihood	Continuous	Pagel (1999)
Blomberg's *K*	Evolutionary	Yes	Permutation	Continuous	Blomberg *et al.* (2003)
D statistic	Evolutionary	Yes	Permutation	Categorical[a]	Fritz and Purvis (2010)
δ statistic	Evolutionary	Yes	Bayesian	Categorical	This work

Note: Diagnostics of phylogenetics signal can be classified based on the approach to estimate phylogenetic signal: evolutionary approaches are generally model-based, while methods that rely on autocorrelation are based on summary statistics of correlation.

[a]Based on the Brownian motion threshold model (Felsenstein, 2005). Table adapted from Münkemüller *et al.* (2012).

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022

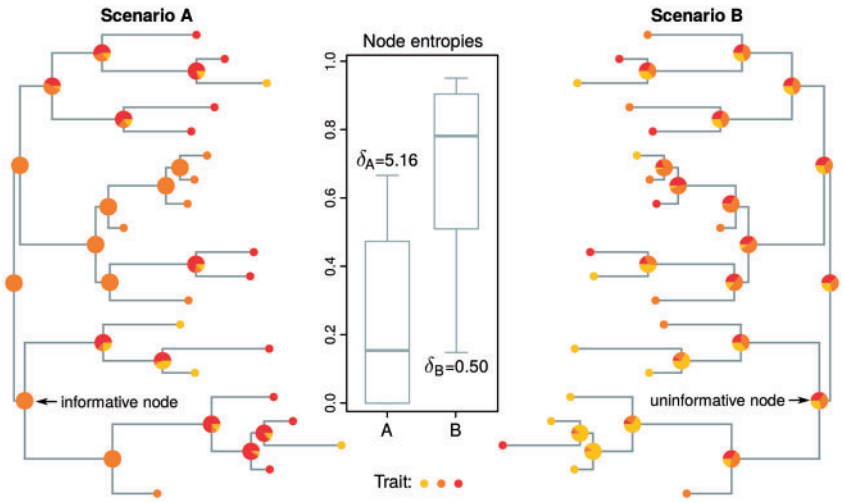

Fig. 1. Node probabilities, entropies and phylogenetic signal. Scenarios A and B depict situations of phylogenetic signal (trait vector with three states follows the hierarchical structure of the phylogeny and therefore closely related species resemble each other more than distantly related ones) and no phylogenetic signal (random trait vector), respectively. Pie charts represent the node probabilities of the ancestral reconstructions for each state. Box plots represent the node entropies for each scenario. Node entropies were calculated based on the linear version of the Shannon entropy presented in this work. In scenario A, more informative inferences are obtained, and therefore node entropies are close to 0. Differently, in situation B, most ancestral inferences are uninformative, which leads to node entropies close to 1. The δ, measuring the degree of phylogenetic signal between a trait vector and a phylogeny, is thus higher in scenario A than B. This example can be reproduced using the code available on the GitHub branch: `mrborges23/delta_statistic`

uncertainty (all the p_i^j are $1/k$), respectively. We summarized our approach for the whole phylogeny using the node entropies of the N nodes and implemented a Bayesian inferential scheme. Given that e^j is defined in the $[0, 1]$ interval, its likelihood was set to follow a beta distribution with parameters α and β. Considering α and β both as having independent exponential priors with rate parameter λ_0, the resulting posterior distribution is (see Supplementary Table S1 for derivation)

$$\log p(\alpha, \beta|e) \propto -Nb(\alpha, \beta) - \alpha\left[\lambda_0 - \sum_j e^j\right] - \beta\left[\lambda_0 - \sum_j \log\left(1 - e^j\right)\right], \quad (3)$$

where $b(.)$ is the logarithm of the Euler integral. To obtain random samples from the posterior distribution, the conditional posteriors of α and β were computed and implemented in a Markov chain Monte Carlo scheme including two Metropolis-Hastings steps in a Gibbs sampler algorithm (Supplementary Table S1). δ is defined as the expected ratio between the posterior distributions of β and α

$$\delta = E\left[\frac{p(\beta|\alpha, e)}{p(\alpha|\beta, e)}\right]. \quad (4)$$

δ is higher than one when $\beta > \alpha$, i.e. when the distribution of entropies favors lower over higher entropies. Hence, the higher the δ-value, the greater the quantity of information given by the ancestral inferences. The code to compute the δ statistic was implemented in the R language and can be accessed in the GitHub branch `mrborges23/delta_statistic`.

2.2 Simulated data

Simulated data was generated for a categorical trait with 2–5 states (k =2, 3, 4 and 5) observed in 10–180 species (s =10, 20, 40, 80 and 160). 1000 random draws of each combination of k and s were obtained using the `rtrait` function (GitHub branch: `mrborges23/delta_statistic`), which simulates the evolution of a categorical trait based on a given binary tree and rate matrix. In particular, we set a time-reversible rate matrix $\boldsymbol{q}$ (Equation 5) in a time-continuous Markov model, defined based on the trait stationary frequencies (π) and the exchangeabilities between the trait states (ρ, with $\rho_{ij} = \rho_{ji}$, where i and j are two possible states)

$$q_{ij} = \pi_i \rho_{ij}. \quad (5)$$

We then examined the behavior of δ for phylogenetic scenarios of interest by decreasing phylogenetic signal, rising trait saturation, increasing tree uncertainty and testing different scenarios of trait evolution: unbiased and biased trait evolution (by respectively setting $\pi \propto 1$ and $\pi \propto (k, k-1, \ldots, 1)$) and trait saturation. These scenarios were simulated using the `rtrait` function, but additional functions from other R packages were also considered (R Core Team, 2015): uncertainty at the topology was employed by the nearest-neighbor-interchange (NNI) using the `rNNI` function in `phangorn` (Schliep, 2011), the matrix exponential was calculated using the `expm` function (Goulet *et al.*, 2017); and the package `ape` allowed to calculate the ancestral probability vectors and generate/manipulate random phylogenetic trees (Paradis *et al.*, 2004).

2.3 Case study

The OrthoMaM database was used to obtain 14 524 maximum likelihood gene trees (Douzery *et al.*, 2014) with up to 43 mammalian species. Overall, 164 gene trees were excluded from the case study for having less than 20 species. Data for the mammalian activity patterns (categorical trait: nocturnal, diurnal and cathemeral; Supplementary Fig. S1) were retrieved from Bennie *et al.* (2014). Ancestral character reconstructions were performed using the `ape` package of the R statistical software using the `ace` function in the `ape` package (Paradis *et al.*, 2004; R Core Team, 2015): i.e. a time-continuous Markov model with an all-rates-different rate matrix. Ancestral probability vectors were obtained by maximum likelihood. Gene ontology (GO) enrichment analysis was performed in the web-based application GOrilla (Eden *et al.*, 2009).

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022

3 Results

δ is proposed to measure the degree of phylogenetic signal between a phylogeny and a categorical variable. We analyzed empirical and simulated data to validate key statistical and phylogenetic aspects of the δ-approach. We have not compared δ with other statistics of phylogenetic signal, as the only one available, the D statistic (Fritz and Purvis, 2010), assumes a Brownian motion threshold model that specifically copes with discrete traits that may have a continuous trait underlying them (Felsenstein, 2005). In our analyses, we do not simulate the evolution of continuous traits, neither we used them to define the categories.

3.1 δ and phylogenetic signal

We tested δ for different scenarios of phylogenetic signal by randomizing half and then for all the species (R50 and R100, respectively, Fig. 2). Our expectation is that δ decreases as trait randomization increases, as ancestral inferences should be more prone to produce unresolved nodes for random or partially random trait vectors. Our results validated this expectation. For all the combinations of the number of species and states tested, δ decays by a factor of 0.25 or less (0.03–0.25, Fig. 2) when the trait vector is completely randomized. We also observed that even when half of the trait vector is randomized δ already significantly decays (0.04–0.34, Fig. 2).

We verified that the loss of phylogenetic signal is more pronounced in the species-rich analysis: δ decreases to less than 15% when 20–160 species are analyzed (Fig. 2). For 10 species, the distribution of δ in R0 and R100 largely overlap, which may complicate the task of attributing δ-values to a scenario of phylogenetic signal or total randomization.

We defined a classification test to pinpoint scenarios of phylogenetic signal or independent evolution. In particular, we fixed the test specificity (or true negative rate) by determining an acceptance threshold that is a quantile in the R0 distribution. R0 is the distribution of δ in scenarios of no phylogenetic signal and therefore permits to calculate a threshold that controls for the percentage of tests that wrongly classify a scenario of independent evolution as evidence of phylogenetic signal. The specificity and sensitivity (or true positive rate) of the test was assessed via ROC (receiver operating characteristic) curves (Fig. 3 and Supplementary Fig. S3). The ROC curves showed that analyses with 10 species have the worst performance: even for considerably high false positive rates (0.10–0.15; Fig. 3), the sensitivity of the test is always lower than 50%. For more than 10 species, we observed that a false positive rate of 0.05 guarantees a good level of sensitivity (higher than 60%; Fig. 3). For more conservative false positive rates (i.e. lower than 0.05), the true positive rate decreases significantly, in some cases, for less than 50% (Fig. 3). Similar results were obtained from the analyses of ROC curves for more than two states (Supplementary Fig. S3).

3.2 δ, trait evolution and saturation

We tested the performance of δ in two possible scenarios of trait evolution: an unbiased scenario, in which all the states are equally preferred (uniform stationary vector $\pi \propto 1$), and a biased scenario, in which one of the states is preferred over the others (asymmetric stationary vector with weights $\pi \propto k : k-1 : \ldots : 1$). We observed that δ is similarly estimated in both scenarios (similar means, medians and quantiles; Fig. 4 and Supplementary Fig. S2B and C).

We tested δ for state saturation by controlling the number of expected state changes in the tree. We define three quantities from low to high number of expected changes: $2k$, $2k + s/2$ and $2k + s$, corresponding to E1, E2 and E3, respectively (Fig. 4). We observed that δ rapidly diminishes as the expected number of state-changes increases (relative differences of 0.01–0.42 for two states, Fig. 4).

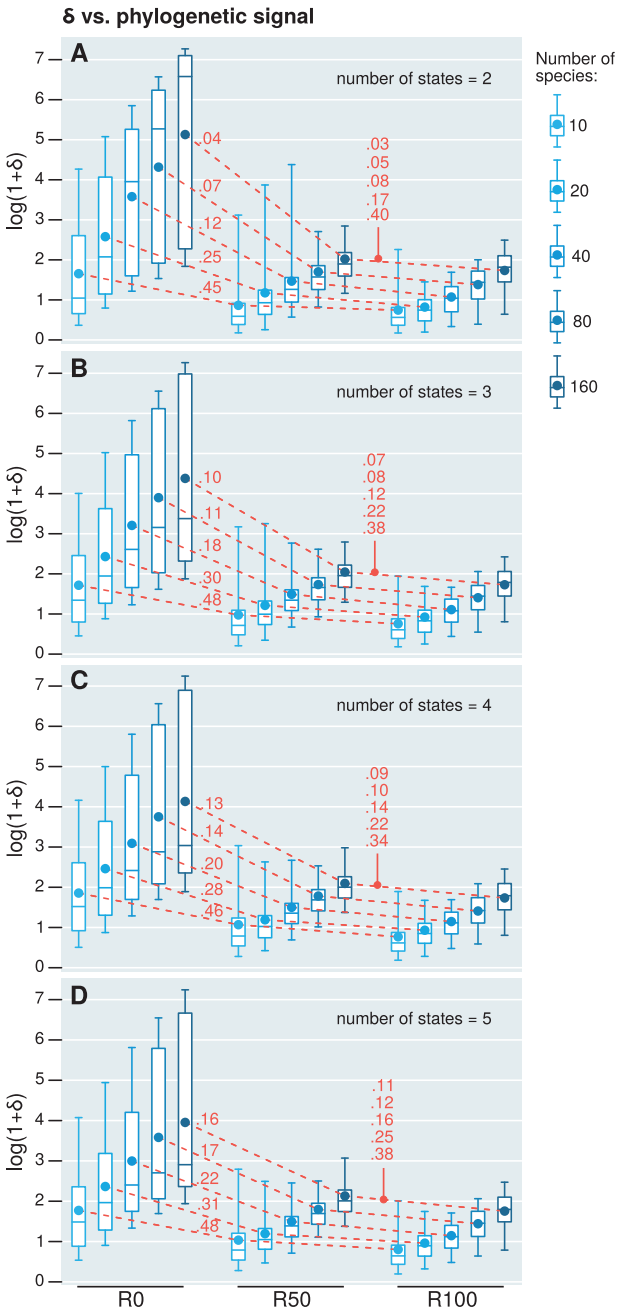

Fig. 2. δ vs. phylogenetic signal. Simulated distributions of δ for scenarios of high (R0), partial (R50) and low (R100) phylogenetic signal. Scenarios of decreased phylogenetic signal were employed by partially or fully randomizing the trait vector. δ-values are represented in the log scale. Each box plot summarizes 1000 simulations while whiskers represent the 5 and 95% quantiles. Gray shaded circles represent the expected value of δ and the dashed lines the relative differences of δ-values (not in the log scale) between the tested scenarios

The decay of δ due to state saturation is always more pronounced for species-rich analyses; differently, δ decays more slowly with state saturation in state-rich analyses (Supplementary Fig. S2B and C).

3.3 δ and tree uncertainty

We tested the influence of branch length and topology uncertainty in δ. We perturbed the branch length by randomizing both half and the total branch lengths (while maintaining tree height; B50 and

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022

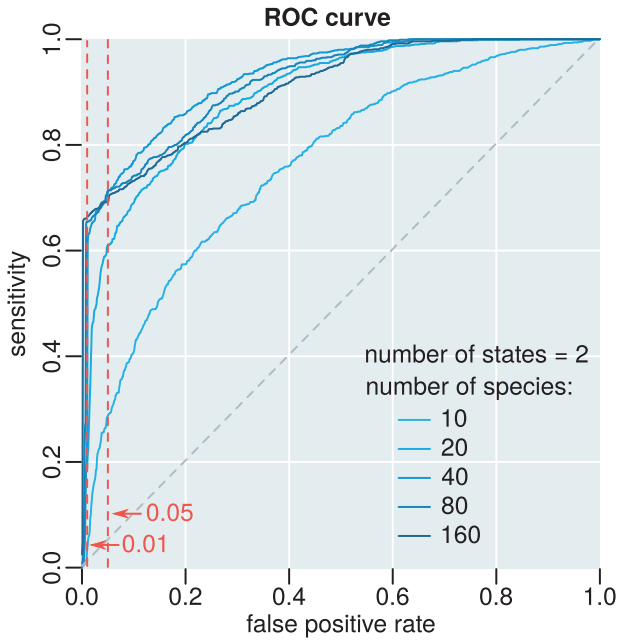

Fig. 3. ROC analysis for classifying δ-values. A classification test was built by fixing the specificity (or the true negative rate, *x*-axis) of the test and defining a δ-threshold that is a quantile in the distribution R0. R0 is the distribution of δ-values in scenarios of independent evolution and therefore permits to calculate the percentage of tests that wrongly classify a scenario of independent evolution as evidence of phylogenetic signal. The sensitivity of the test (or the true positive rate, *y*-axis) corresponds to the percentage of tests that correctly classify a scenario of phylogenetic signal. ROC curves for three to five states are in Supplementary Figure S3

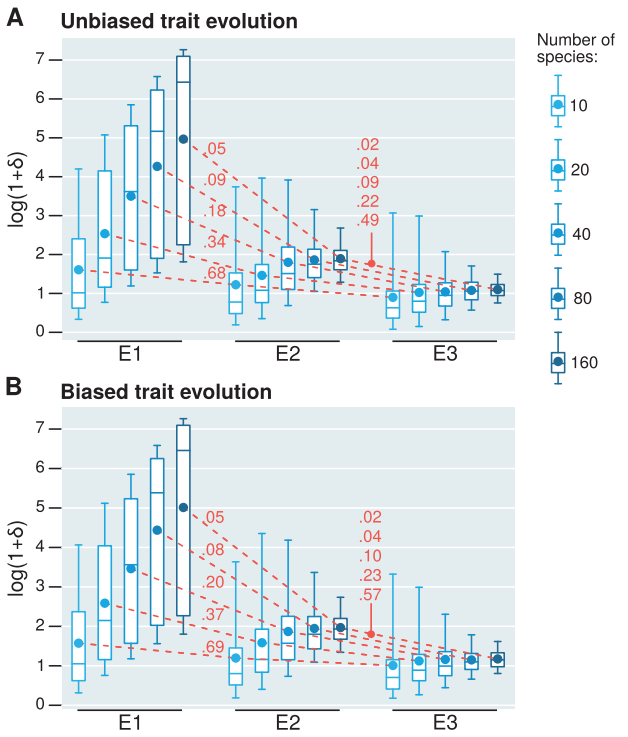

Fig. 4. δ vs. type and rate of evolution. Distributions of δ for scenarios of unbiased (**A**) and biased (**B**) trait evolution were simulated by setting uniform and asymmetric stationary vectors, respectively. δ-values are represented in the log scale. The impact of trait saturation on δ is simulated by setting an increasing number of expected character-changes per tree: E1, E2 and E3 refer to $2k$, $2k + s/2$ and $2k + s$ events, respectively. Each box plot summarizes 1000 simulations while whiskers represent the 5 and 95% quantiles. Gray shaded circles represent the expected value of δ and the dashed lines the relative differences of δ-values (not in the log scale) between the tested scenarios. The plots for three to five states are in Supplementary Figure S2B and C

B100, respectively; Fig. 5A). Perturbations on the topology were employed by the NNI tree rearrangement affecting 10 and 50% of the branches of the true topology (NNI10 and NNI50, respectively; Fig. 5B). We observed that uncertainty at the topology affects the distribution of δ more than the branch length uncertainty (Fig. 5): the relative differences of δ-values were of 0.4–0.87 and 0.31–0.58 in BL50 and NNI50 (for $k = 2$, Fig. 5A and B), despite the same number of branches being perturbed in each case. Similar patterns were observed for three to five states, but with a slightly less pronounced effect of branch length and topology uncertainty (i.e. δ decays slowly, Supplementary Fig. S2D and E, $k = 3$–5). We further observed that the higher the number of species in the analysis, the more pronounced the effect of branch length and topology uncertainty on δ: an expected impact on δ of -31 to -11% was obtained for 160 species, while an expected difference of -4 to -11% was obtained for 10 species; Fig. 5 and Supplementary Fig. S2D and E.

3.4 Selecting genes associated with mammalian lifestyles

Mammals show a great variety of activity patterns: while most mammals are diurnal or nocturnal, some can also be cathemeral (characterized by the equal use of the night and day) (Borges *et al.*, 2018; Walls, 1942). We employed the δ-statistic in 14 360 case study gene trees to unravel those associated with the mammalian lifestyles (three categories: nocturnal, diurnal and cathemeral). In addition, we calculated the distribution of δ for scenarios of random trait evolution: we simulated random trees and random trait-vectors with three characters for 37 species (average number of species per tree in the case study; trees with less than 20 species were excluded from the analysis). We observed that the distribution of the case study δ clearly deviates from the simulated one (medians of 2.4 and 0.7, respectively; Fig. 6A), with the case study δ-values being on average higher than the random δ-values. We obtained 9400 genes showing evidence of phylogenetic signal with mammalian activity patterns (δ threshold of 2.1, by defining a test specificity of 95%; Fig. 6A).

To check whether our results have biological signatures of interest, we performed enrichment analyses for GO terms, by ranking the study genes according to their δ-values (i.e. top-ranked genes have higher δ). We found three significant GO categories (GO: 0050907, GO: 0050911 and GO: 0009593, FDR corrected *P*-values = 0.007, 0.016 and 0.021, respectively; Supplementary Table S2) appearing densely at the top-ranked genes. These GO-terms are all associated with the detection of chemical stimulus involved in sensory perception of smell. Selected genes include mostly olfactory receptors, but also taste receptors and receptor transporter/interacting proteins (Supplementary Table S2). We observed that olfactory receptors have generally an elevated δ-value, being most of them above the δ threshold for evidence of phylogenetic signal (Fig. 6B).

4 Discussion

The use of δ can be generalized to every model or type of inference that returns probability vectors for the ancestral nodes (which is not the case for parsimony); these are required to calculate the node

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022

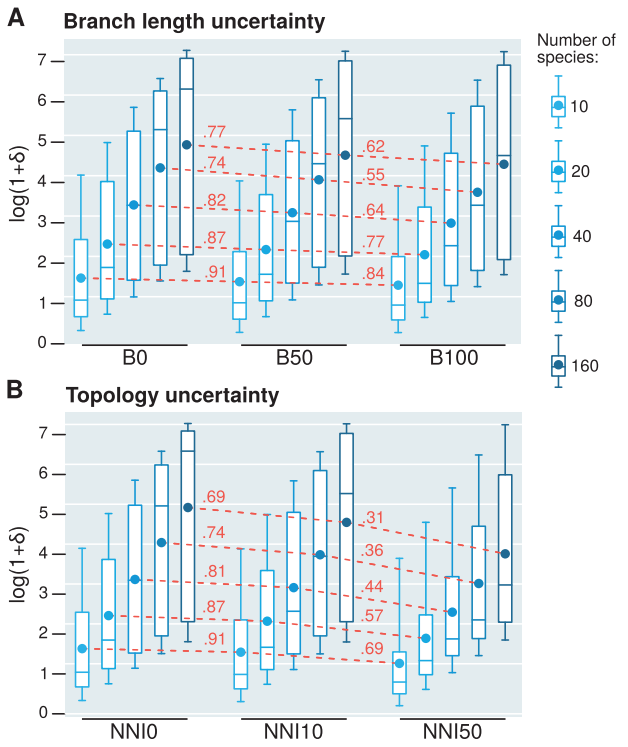

Fig. 5. δ vs. tree uncertainty. (**A**) Simulated distributions of δ for scenarios of branch length uncertainty: 50% (B50) and 100% (B100) of the true branch lengths (B0) were re-sampled. (**B**) Simulated distributions of δ for scenarios of topology uncertainty: perturbations on the true topology (NNI0) were employed by performing a NNI rearrangement on 10% (NNI10) and 50% (NNI50) of the tree branches. δ-values are represented in the log scale. Each box plot summarizes 1000 simulations, while whiskers represent the 5 and 95% quantiles. Gray shaded circles represent the expected value of δ and the dashed lines the relative differences of δ-values (not in the log scale) between the tested scenarios. The plots for three to five states are in Supplementary Figure S2D and E

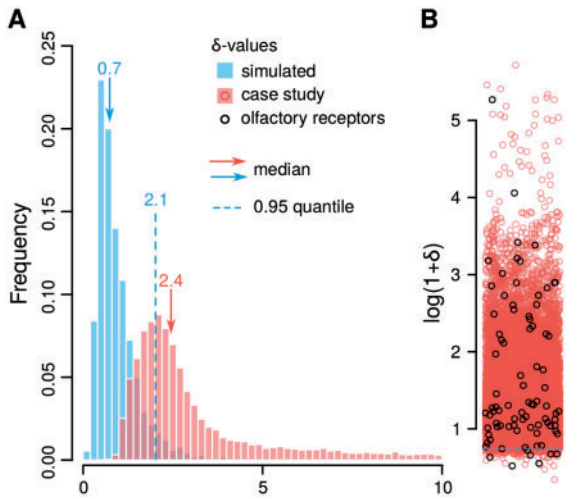

Fig. 6. δ-values for the case study 14 360 gene trees. (**A**) Distribution of the δ-values for simulated (light gray) and the case study data (dark gray). Simulated δ-values were obtained by recreating scenarios of no phylogenetic signal (simulation conditions: three states and 37 species). (**B**) Distribution of the case study δ on the log scale, depicting the olfactory receptor genes (black dots)

entropies. Here, we used a time-continuous discrete-state Markov chain to obtain the ancestral inferences, but δ could be easily integrated with more-complex models. These include the more-complex semi-Markovian and the non-stationary Markov models that are becoming increasingly common in phylogenetics (Kaehler *et al.*, 2015; McAuliffe *et al.*, 2004).

4.1 δ: a measure of phylogenetic signal

We validated δ as a possible approach to measuring the degree of phylogenetic signal between categorical traits and phylogenies. We showed that δ increases when the trait evolves according to the phylogeny, but more importantly, that δ decreases when the trait evolves independently. The δ can be any positive real number: the higher the δ-value, the higher the degree of phylogenetic signal between a given trait and the phylogeny.

We tested the statistical behavior of δ for several conditions of the number of species and trait states in the analysis. We proposed a test that classifies δ-values based on the distribution of δ given no phylogenetic signal (i.e. random attribution of trait vector to the phylogeny). This distribution can be easily simulated for the number of species and states in the analysis. Our results showed that fixing the specificity of the test to 95% guarantees a good true positive rate (sensitivity higher than 60%). Our test is conservative in the sense that it may reject some δ-values of phylogenetic association; however, that is the best compromise to avoid wrongly classifying δ-values as evidence of phylogenetic signal (which may happen for more stringent false positive rates).

Our results showed that the number of species is a key aspect to consider when analyzing δ. We observed that the variance of δ is very large for 10 species, which made it difficult to differentiate δ from a scenario of no phylogenetic signal. In addition, the sensitivity of the test is always very low for 10 species ($\ll$50%), which will most likely produce false negatives. Therefore, we recommend using δ when 20 or more species are analyzed.

4.2 δ can be widely applied in phylogenetic contexts

We verify that the performance of δ does not seem to significantly decrease if there is some degree of tree uncertainty. However, uncertain topologies are more prone to affect δ than branch length uncertainty. Uncertainty regarding the topology can be assessed through the posterior clade probabilities or bootstraps (Guindon *et al.*, 2010; Huelsenbeck and Ronquist, 2001). We verified that topology movements affecting 10% of the branches of the true topology may decrease δ; however, the distribution of δ-values is still far away from the situation of no phylogenetic signal. Thus, in principle, we may use the δ approach even when some of the nodes in the tree have low posterior clade probability or bootstrap values.

Branch length uncertainty is more difficult to assess because, unlike for topology, there are no statistics to assess this. We may use the branch length variance as a proxy for branch length uncertainty, and testing the impact of changing those branch lengths on δ. The coefficient of variation (also known as the relative standard deviation) can be employed to determine the less precise branch lengths. In principle, if the topology is well supported, branch length uncertainty should not be a major issue. Indeed, despite δ decreasing when 50% of branches are perturbed, the distribution of δ does not really look like the scenario of no phylogenetic signal. Thus, while we may expect our test to have less power, we can still implement δ when some branches have high variance. The long-branch attraction [i.e. when distantly related lineages are incorrectly inferred to be closely related (Bergsten, 2005)] is an issue that may significantly

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022

affect δ as it disturbs both the branch lengths and the topology. Therefore, we call special attention when using the δ-approach to measure phylogenetic signal in phylogenetic trees affected by the long-branch attraction artifact.

The performance of δ is maintained regardless of the trait evolving in an unbiased or biased way. These scenarios express situations in which the trait evolves under neutral and directional selection, respectively. It seems that as long as the trait follows the tree's hierarchical structure, the δ-approach is able to retrace the evolution of the trait and, thus, recognize phylogenetic signal between the tree and the trait. Another possible evolutionary scenario is when the trait evolves directionally in some clades, but neutrally in the others: we have not tested this situation but our results suggest that δ would perform equally well as long as the rate of trait evolution is approximately equal in both cases. Differently, in chimeric scenarios, in which both selection and drift act with very different rates, we may expect a decreasing of the δ-statistic as it should be more difficult to retrace trait's evolution.

Another aspect that we have observed is that fast-evolving traits may mimic scenarios of no phylogenetic signal. Indeed, if the trait evolves in such a way that it re-evolves multiple times (including several reversible changes), then it may happen that distantly related data may resemble more than closely related data (Rheindt *et al.*, 2004). Consequently, one should expect δ to decrease with trait saturation. It is important to notice that phylogenetic signal mistakes fast-evolving traits with independent evolution, therefore, care must be taken when assuming that lack of phylogenetic signal implies independent evolution.

4.3 δ captures biological signatures of phenotypic evolution

We implemented the δ-statistic as a proxy to unravel the protein-coding genes co-evolving with the mammalian lifestyles. Our case study permitted us to identify 9400 gene trees evolving with a considerable degree of phylogenetic signal with the mammalian activity patterns. Among these genes, we found several olfactory receptors with high δ-values, and we further validated the olfactory reception as being deeply associated with the activity pattern in mammals. Modifications of the sensory system are among the most common changes that occur when shifting from an activity pattern to another (Barton *et al.*, 1995). It is therefore expected that several olfactory receptors were pinpointed by our analysis. Several works confirm the results obtained here, as several olfactory-related adaptations have been linked to the evolution of the activity patterns in mammals. For example, studies confirming the relationship among olfactory receptor genes expansions and retractions, divergence and function with the mammalian lifestyles can be found in Kishida (2008), Wang *et al.* (2010), Hayden *et al.* (2010), Khan *et al.* (2015), Tsagkogeorga *et al.* (2017) and Hughes *et al.* (2018). Thus, our statistic seems to be a valuable tool to pinpoint genes mediating the evolution of phenotypes.

On a final note, the δ-approach can be easily adapted to any other categorical trait one may want to study. Indeed, δ can be easily integrated with other designs, namely by considering: different evolutionary models of trait evolution; species trees instead of gene trees; diverse evolutionary distances (e.g. $\omega = dN/dS$); and alternative inferential methods to obtain node probabilities.

5 Conclusion

Here, we developed the δ statistic that uses the concept of entropy to test phylogenetic signal between gene trees and categorical traits. The δ can be simultaneously useful to finding genes that co-evolved with traits of interest, as well as understanding the evolution of complex traits (i.e. those requiring the integrated evolution of several genetic players). Therefore, δ is a useful approach to unravel the molecular basis of those phenotypes that can only be characterized in categories.

Acknowledgements

We thank James Howie and Juraj Bergman for helpful comments on the manuscript. We thank the anonymous reviewers for their many insightful comments and suggestions.

Funding

RB and CG were funded with a PhD grant from Fundação para a Ciência e a Tecnologia (Foundation for Science and Technology, FCT) (RB: SFRH/BD/79850/2011 and CG: SFRH/BD/71041/2010). APR was partially supported by CMUP (UID/MAT/00144/2013), which is funded by FCT with national (MEC) and European structural funds through the program FEDER, under PT2020 and project STRIDE—NORTE-01-0145-FEDER-000033 funded by ERDF—NORTE 2020. AA was partially supported by the Strategic Funding UID/Multi/04423/2013 through national funds provided by FCT and the European Regional Development Fund (ERDF) in the framework of the program PT2020, by the European Structural and Investment Funds (ESIF) through the Competitiveness and Internationalization Operational Program - COMPETE 2020 and by National Funds through the FCT under the project PTDC/AAG-GLO/6887/2014 (POCI-01-0124-FEDER-016845).

Conflict of Interest: none declared.

References

Abouheif,E. (1999) A method for testing the assumption of phylogenetic independence in comparative data. *Evol. Ecol. Res.*, **1**, 895–909.

Antunes,A. *et al.* (2008) The evolutionary dynamics of the lion panthera leo revealed by host and viral population genomics. *PLoS Genet.*, **4**, e1000251.

Barton,R.A. *et al.* (1995) Evolutionary radiation of visual and olfactory brain systems in primates, bats and insectivores. *Philos. Trans. R. Soc. B: Biol. Sci.*, **348**, 381–392.

Bennie,J.J. *et al.* (2014) Biogeography of time partitioning in mammals. *Proc. Natl. Acad. Sci. USA*, **111**, 13727–13732.

Bergsten,J. (2005) A review of long-branch attraction. *Cladistics*, **21**, 163–193.

Blomberg,S.P. *et al.* (2003) Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, **57**, 717–745.

Borges,R. *et al.* (2018) Adaptive genomic evolution of opsins reveals that early mammals flourished in nocturnal environments. *BMC Genomics*, **19**, 121.

Burleigh,J.G. and Mathews,S. (2004) Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *Am. J. Bot.*, **91**, 1599–1613.

Douzery,E.J.P. *et al.* (2014) OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol. Biol. Evol.*, **31**, 1923–1928.

Dunn,M. (2005) Structural phylogenetics and the reconstruction of ancient language history. *Science*, **309**, 2072–2075.

Eden,E. *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Felsenstein,J. (2005) Using the quantitative genetic threshold model for inferences between and within species. *Philos. Trans. R. Soc. B Biol. Sci.*, **360**, 1427–1434.

Fritz,S.A. and Purvis,A. (2010) Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.*, **24**, 1042–1051.

Gilbert,G.S. and Webb,C.O. (2007) Phylogenetic signal in plant pathogen-host range. *Proc. Natl. Acad. Sci. USA*, **104**, 4979–4983.

Goulet,V. *et al.* (2017) *expm: Matrix Exponential, log, 'etc'*. R Package Version, 0.999-2.

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022

Guindon,S. *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.

Hayden,S. *et al.* (2010) Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res.*, **20**, 1–9.

Huelsenbeck,J.P. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Hughes,G.M. *et al.* (2018) The birth and death of olfactory receptor gene families in mammalian niche adaptation. *Mol. Biol. Evol.*, **35**, 1390–1406.

Kaehler,B.D. *et al.* (2015) Genetic distance for a general non-stationary Markov substitution process. *Syst. Biol.*, **64**, 281–293.

Kamilar,J.M. and Cooper,N. (2013) Phylogenetic signal in primate behaviour, ecology and life history. *Philos. Trans. R. Soc. B Biol. Sci.*, **368**, 20120341–20120341.

Khan,I. *et al.* (2015) Olfactory receptor subgenomes linked with Broad Ecological Adaptations in Sauropsida. *Mol. Biol. Evol.*, **32**, 2832–2843.

Kishida,T. (2008) Pattern of the divergence of olfactory receptor genes during tetrapod evolution. *PLoS One*, **3**, e2385.

Losos,J.B. (2008) Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecol. Lett.*, **11**, 995–1003.

McAuliffe,J.D. *et al.* (2004) Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics*, **20**, 1850–1860.

Moran,P.A.P. (1950) Notes in continuous stochastic phenomena. *Biometrika*, **37**, 17–23.

Münkemüller,T. *et al.* (2012) How to measure and test phylogenetic signal. *Methods Ecol. Evol.*, **3**, 743–756.

Pagel,M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.

Paradis,E. *et al.* (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Pearman,P.B. *et al.* (2008) Niche dynamics in space and time. *Trends Ecol. Evol.*, **23**, 149–158.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rheindt,F.E. *et al.* (2004) Rapidly evolving traits and the comparative method: how important is testing for phylogenetic signal? *Evol. Ecol. Res.*, **6**, 377–396.

Schliep,K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.

Simmons,M.P. and Ochoterena,H. (2000) Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.*, **49**, 369–381.

Thuiller,W. *et al.* (2011) Consequences of climate change on the tree of life in Europe. *Nature*, **470**, 531–534.

Tsagkogeorga,G. *et al.* (2017) Comparative genomics reveals contraction in olfactory receptor genes in bats. *Sci. Rep.*, **7**, 259.

Walls,G.L. (1942) *The Vertebrate Eye and Its Adaptive Radiation*. Cranbrook Institute of Science, Bloomfield Hills, MI.

Wang,G. *et al.* (2010) Comparative genomic analysis reveals more functional nasal chemoreceptors in nocturnal mammals than in diurnal mammals. *Chin. Sci. Bull.*, **55**, 3901–3910.

Downloaded from https://academic.oup.com/bioinformatics/article/35/11/1862/5144670 by guest on 16 August 2022