Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 251 (2008) 606–615
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Evolutionarily distinctive species often capture more phylogenetic diversity than expected

David W. Redding[a,b,*,1], Klaas Hartmann[c,1], Aki Mimoto[b,1], Drago Bokal[b,d],
Matt DeVos[b,e], Arne Ø. Mooers[a,b,f]

[a]*BISC, Simon Fraser University, Burnaby, Canada V5A 1S6*
[b]*IRMACS, Simon Fraser University, Burnaby, Canada V5A 1S6*
[c]*Biomathematics Research Center, University of Canterbury, Christchurch, New Zealand*
[d]*Institute of Mathematics, Physics, and Mechanics, Ljubljana, Slovenia*
[e]*Department of Mathematics, Simon Fraser University, Burnaby, Canada*
[f]*Institute for Advanced Study, Berlin, Germany*

Received 28 June 2007; received in revised form 26 November 2007; accepted 11 December 2007
Available online 4 March 2008

Abstract

Evolutionary distinctiveness measures of how evolutionarily isolated a species is relative to other members of its clade. Recently, distinctiveness metrics that explicitly incorporate time have been proposed for conservation prioritization. However, we found that such measures differ qualitatively in how well they capture the total amount of evolution (termed phylogenetic diversity, or PD) represented by a set of species. We used simulation and simple graph theory to explore this relationship with reference to phylogenetic tree shape. Overall, the distinctiveness measures capture more PD on more unbalanced trees and on trees with many splits near the present. The rank order of performance was robust across tree shapes, with apportioning measures performing best and node-based measures performing worst. A sample of 50 ultrametric trees from the literature showed the same patterns. Taken together, this suggests that distinctiveness metrics may be a useful addition to other measures of value for conservation prioritization of species. The simplest measure, the age of a species, performed surprisingly well, suggesting that new measures that focus on tree shape near the tips may provide a transparent alternative to more complicated full-tree approaches.
© 2008 Elsevier Ltd. All rights reserved.

Keywords: Conservation priorities; Distinctiveness; Evolutionary value; Genetic uniqueness

---

1. Introduction

With increasing extinction there is a pressing need to effectively prioritize species for conservation. Many non-exclusive currencies are used, e.g. threat status, ecological importance, social or intrinsic value, and financial cost (for discussion, see Crozier, 1992; Weitzman, 1998; Andelman, 2004; Avise, 2005). Here we focus on the evolutionary distinctiveness of species in the context of their conservation. In particular, we examine the trade-off between prioritizing the most evolutionary distinctive species in a tree and prioritizing sets of species that best represent the whole tree.

Conservation biologists have approached the goal of representing a phylogenetic tree from two angles. Both approaches use information about the relatedness among tips (usually species), but one (phylogenetic diversity, PD) is a group measure, while the other (evolutionary distinctiveness) is a species-specific property. To illustrate the connections between the two approaches, consider the order *Sphenodontia*. This order contains the two species of tuatara and is sister to *Squamata* (Snakes, Lizards, Amphisbaenians), which contains ~6200 species. From a macroevolutionary perspective, if one species from each

---

*Corresponding author at: BISC, Simon Fraser University Burnaby, Canada V5A 1S6. Tel.: + 1 604 268 7072.

E-mail addresses: dredding@sfu.ca (D.W. Redding), amooers@sfu.ca (A.Ø. Mooers).

[1] These authors contributed equally.

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2007.12.006

order were equally threatened, priority should go to a tuatara before any lizard, snake, or amphisbean species, because both tuatara species are highly distinctive and so contain a disproportionately large proportion of the PD contained within the two orders. However, if only two species of the ~6202 species were to be preserved (an unlikely scenario), the tree would be best represented with a set that included only one of the two tuataras, and one Squamate. More generally, any *subset* that did not contain *one* of the tuataras would be suboptimal.

The idea of comparing the relative PD represented by *sets* of species in order to prioritize sets that contribute more unique evolution was pioneered by Vane-Wright et al. (1991) and by Faith (1992). The PD of a set of species is generally measured as the sum of the branch lengths of the tree containing those species and the root (see Fig. 1 and Faith and Baker, 2006). Sets of species with maximal PD can be found using simple algorithms (Steel, 2005; Pardi and Goldman, 2005; Minh et al., 2006), and the approach has also been useful when ranking predefined areas (which define subsets) for conservation (see, e.g. Forest et al., 2007). The PD approach has also been extended to include species survival probabilities and conservation costs and budgets (Weitzman, 1998; Hartmann and Steel, 2006, 2007; Pardi and Goldman, 2007).

In parallel, systematists have proposed metrics for how much unique evolution a particular species contributes to some larger set (again, see Vane-Wright et al., 1991; see also May, 1990; Nixon and Wheeler, 1992; Pavoine et al., 2005; Redding, 2003; Redding and Mooers, 2006; Isaac et al., 2007). Early attempts to attribute a score of evolutionary distinctiveness to individual species (May, 1990; Nixon and Wheeler, 1992; Vane-Wright et al., 1991) used only tree topology, and relied on the fact that basal and evolutionary isolated species have fewer nodes between the tip and the root. Recent workers (Isaac et al., 2007; Pavoine et al., 2005; Redding and Mooers, 2006; Steel et al., 2007; Weitzman, 1998) have suggested distinctiveness measures (outlined below) that use both topology and internal branch lengths to measure species isolation. All such measures have one thing in common: they give species that have many and closer relatives less value than they give species with fewer and more distant relatives.

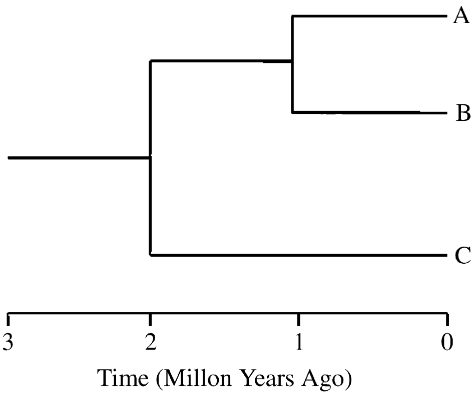

Fig. 1. This figure shows a tree connecting a hypothetical group of species. The phylogenetic diversity (PD) of all the species is found by summing up the branch lengths in the tree, here 6 million years. The PD of a subset of species is found by summing up the branch lengths of the tree connecting those species and the root, for group AB it is 4 million years. In this paper we investigate how well species prioritizations based on simple indices capture the PD of the tree.

Evolutionary distinctiveness measures and PD approaches differ in substantial ways. PD is uninformative for any one species on an ultrametric tree—all single species are the same distance from the root and so receive the same value. Many current conservation approaches (e.g. endangered species lists) rely on having species ranked in order of priority. Current PD approaches offer no such order. To overcome this, species within any optimal set chosen could be ordered by arranging them according to their evolutionary distinctiveness, or alternatively, species could be chosen according to natural species-specific indices that ensure optimal sets (DB, MdV, KH, unpublished results). However, there will be as many possible rankings of species produced by these approaches as there are PD maximizing solutions for the set of candidate species, and thus may be difficult to implement at the management level.

More importantly, the amount of PD saved is only optimal if *all* the species that are selected are subsequently protected. If any species in the selection are lost, new optimal sets are possible. Finally, it may be difficult to find optimal sets of species if there are large numbers of species to prioritize, and other complex factors such as cost of conserving individual species are considered (Weitzman, 1998).

The recently developed species-specific measures of evolutionary distinctiveness may, in comparison, represent a flexible and transparent conservation tool to promote “evolutionary value” in the current legislative climate. However, and importantly, they have not been designed to capture total PD. If sets of evolutionary distinctive species did capture substantial PD, then the species-specific measures would be doubly useful, highlighting the most individually distinctive species and helping preserve more of the tree of life. This is the focus of the present study. We first outline the distinctiveness measures we tested, and then consider aspects of the underlying tree that might affect the relationship between distinctiveness measures and PD.

2. The measures

There exists a small group of older, node-based measures of distinctiveness (reviewed in Pavoine et al., 2005). This group suffers from poor resolution among species (i.e. it gives rise to many tied values) and also loses any information conferred by branch lengths (Pavoine et al., 2005). However, preliminary analyses (results not shown) identified May’s (1990) modification of Vane-Wright et al.’s (1991) measure (VW) as the best node-based measure for capturing PD, and so we used it as a basis for comparison:

$$VW(T,i) = \frac{1}{\sum_{v \in q(T,i,r)} deg_{out}(v)}, \tag{1}$$

where for a tree T, $q(T,i,r)$ denotes the set of vertices between species i and the root, and $deg_{out}(v)$ denotes the number of lineages emanating from vertex v (two for a bifurcating tree).

Quadratic entropy (QE; Rao, 1982) was initially proposed as a measure of the biodiversity for a species assemblage. This measure considers both relative abundance and relatedness and returns the expected phylogenetic distance between two randomly chosen individuals. Pavoine et al. (2005) observed that a solution that assigns idealized individual species proportions to yield the maximal diversity value satisfies the criteria of an distinctiveness measure: if species-pairs are sampled in proportion to their distinctiveness, pairwise phylogenetic distance is maximized. So, Pavoine et al. (2005) assign weights (p_i) that maximize $QE(T,p)$:

$$QE(T,p) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(p_i p_j \sum_{k \in s(T,i,j)} \lambda_k \right). \tag{2}$$

Here $s(T,i,j)$ is the set of edges between species i and j, and λ_k is the length of the kth edge. The weights (p_i) that maximize this quantity are the indices assigned to the species. We draw attention to the fact that QE is a pairwise measure.

Equal splits (ES), on the other hand, simply distributes the phylogenic diversity contained within the tree uniquely among the species at the tips, achieved by dividing the phylogenetic distance represented by a branch equally among its daughter branches. The sum of the ES value from every internal branch between a species tip and the root is a representation of the unique amount of evolutionary history a species embodies. This process gives species from species-poor clades a higher weighting as less of the path length between the tip and the root is shared by other species. Sharing the entire branch lengths among daughter branches ensures that sum of the ES measure across the tips equals the total PD of the tree (Pauplin, 2000; Semple and Steel, 2003):

$$ES(T,i) = \sum_{e \in s(T,i,r)} \left(\lambda_e \prod_{f \in q(T,i,e)} \frac{1}{deg_{out}(f)} \right), \tag{3}$$

where for convenience we take $q(T,i,e)$ to be the set of vertices between the species and the bottom of edge e. ES for species i is the sum of all edge lengths λ_e between the species and the root, with each edge length being divided by the product of all the outward degrees of vertices between the edge and the species.

A conceptually similar metric distributes internal branches fairly, not to descendant clades as in ES, but to all descendant species (Isaac et al., 2007; Redding, 2003):

$$FP(T,i) = \sum_{e \in s(T,i,r)}^{r} \frac{\lambda_e}{c_e}. \tag{4}$$

Fair proportion (FP) for species i is the sum of all edge lengths λ_e between the species and the root, with each edge length being divided by the number of species c_e in the clade it subtends. This metric has been used in a recently suggested prioritization framework (but called "ED"; Isaac et al., 2007).

Unlike the first four measures, the age of the species or its pendant edge (PE) value (Altschul and Lipman, 1990) disregards the internal structure within a tree. PE is the distance from any tip on a tree to where it subtends the tree of life (for example, the PE value for species A in Fig. 1 is 1 million years). This gives it the advantage of being an absolute measure. A PE value can only change if the phylogenetic relationships of sister species are revised or re-dated. Denoting the set of all species in T by X, the PE value is expressed mathematically as

$$PE(T,i) = PD(X) - PD(X - \{i\}). \tag{5}$$

Finally, for comparison, we included the expected value of PD if species were chosen at random (AVG, following Nee and May, 1997, using a calculation presented in our Appendix).

3. Tree shape

Tree shape is likely to be an important factor in determining how effectively distinctiveness measures capture PD. We outline three measures of tree shape and how we think they will affect the PD 'capture rate'.

The balance of the tree towards the root (I_c; Colless, 1982; Heard, 1996) will dictate whether, when randomly selecting species, some internal branches are more likely to be chosen than others. Repeatedly selecting closely related species will decrease the total amount of PD represented, since the same internal branches are chosen again and again. Random selection on trees with imbalance at the root will have this effect.

The tree shape measure I_2 (Mooers and Heard, 1997) measures the imbalance over the entire tree (Matsen, 2006). In balanced trees or areas of the tree, tips will have similar distinctiveness scores (e.g. there will be the same number of nodes between the tips and the root). Distinctiveness measures will then rank these species similarly, and will choose closely related species, decreasing the total PD captured. On small trees, I_2 is strongly correlated with the number of terminal pairs, or cherries, there are in a tree (for 16-tip Yule trees, this relationship is very strong: Pearson's Rho $= 0.988$, $n = 5005$, $p < 0.001$). Cherries will have the same distinctiveness measure. On larger trees, there are more complex sub-tree shapes that can act like cherries. Consider a large tree with high imbalance at the root but balanced elsewhere. The species on the species-poor side of the tree are likely to have fewer edges separating them from the root relative to the species on the other side of the tree. A simple distinctiveness measure such as VW (see Eq. (1)) is likely to choose species exclusively from the species-poor side of the tree without crossing the root, decreasing the amount of PD captured.

The final aspect of a tree branching structure that is likely to affect capture rate is average node depth, which can be measured with gamma (Pybus and Harvey, 2000), as it approximates how much of the tree there is to "share" among the tips. Consider a star phylogeny (gamma $\gg 0$), where species contribute equally to PD value ($\sim PD/n$): each species contributes the same amount to total PD; at the other extreme, where all but one of the nodes are at the present (gamma $\gg 0$), the first species chosen captures $\sim PD/2$, and most additional species contribute little. In these extreme cases, the capture rates are fixed and irrelevant of the order in which species are chosen and all the measures, therefore, must perform similarly.

Trees with gamma values between these two extremes are expected to have much more complicated PD capture curves, but must be bounded by these two examples. In this middle range of gamma, there are many different possible tree topologies, with, therefore, much greater variation in the proportion of the tree a tip can represent, and as a result, tip choice must be an important factor for capturing PD.

To test how efficiently distinctiveness metrics capture PD, and to explore how tree shape might affect this efficiency, we undertook simple simulations, and then formalized some of the results using a graph-theoretical approach, and finally applied the metrics to a sample of trees derived from empirical data.

4. Methods

Our primary data set consisted of 5000 simulated "Yule trees" (Yule, 1924) with 16 tips, created using Bio::Phylo package (Vos, 2006). In order to test the sensitivity of the findings to the process model used, we also simulated "Hey trees" (Hey, 1992): these are the tree shapes expected under the Moran (1951) coalescent and have Yule topologies, but different waiting times, with more splits occurring near the present. Using Yule trees in this situation is conservative, as the principal way they differ to "real trees" is their limited imbalance (Mooers and Heard, 1997). Thus, if as predicted, the greater imbalance leads to better performance by distinctiveness measures over random choice, samples of real trees should show a stronger relationship.

We confine ourselves to ultrametric trees. This allows us to set a constant "currency", time. We note that many of the arguments we present apply to trees with different currencies, such as trait richness, but that QE is only applicable to ultrametic trees. Though we lose generality by confining ourselves to ultrametric trees, at present we can think of no measure besides time that can be as widely applied across the Tree of Life.

For each simulated 16-tipped tree, we used the five metrics to rank the 16 tips (ES, FP, QE, VW, and PE). We then recorded the cumulative PD captured (as a proportion of that represented by the entire tree) when selecting 2, 4, 6, 8, and 10 species. When ranks of species tied, we took the mean (expected) value of PD (i.e. that if the choice among tied species were random; see Appendix).

We first asked if we could predict the amount of PD captured (by the five distinctiveness measures and the AVG algorithm) using a series of exploratory logistic regression models, with the y variable being percentage of the tree captured, and our three measures of tree shape as x variables: I_c, I_2, and gamma. In total, five separate models were constructed with the PD captured by 2, 4, 6, 8, and 10 species as the respective dependent variables. For statistical independence, we recorded the proportion of PD captured by only one of six measures (the five distinctiveness metrics, and the average (expected) proportion) on each of the 5000 trees, yielding $n = 833$ in our models.

We selected models in a step-wise manner using the conservative Bayesian Information Criteria (BIC) score as the selection criterion (in R environment; *step* function, Venables and Ripley, 2002). Because highly unbalanced tree topologies are highly constrained, model residuals were non-normally distributed. We, therefore, used bootstrapping (*lmboot* in R, Peng, 2005) to estimate the error residuals about the coefficients.

Finally, we applied the five distinctiveness metrics to a selection of 50 "real trees", representing a large range of tip number (4–89 tips) and a diverse taxonomic cross-section (including flagellate protests, dicotyledons, birds, and mammals; McPeek and Brown, 2007) and asked how much PD was represented by the most original set of species at two different group sizes: 25% and 50% of total species number, using the same methods as for the simulated data set. We then compared these values to the expected PD when randomly choosing groups of the same sizes (i.e. 25% and 50% of the species) from these trees.

5. Results

For our small simulated trees, all five distinctiveness metrics perform well at capturing PD for the majority of tree shapes, and generally capture significantly more of the tree than would a random sample (Table 1). The extra amount of the tree that would be captured by selecting species using an distinctiveness measure as opposed to randomly selecting them, e.g. the difference between random selection and ES, ranges from 4% to 9%.

When two species are chosen from the trees, the measures capture between 28% (VW) and 30% (ES) of the total PD, with all five measures doing better than random in this regard; corresponding numbers for eight species are 72–77% (Table 1). The ES measure captured significantly more PD than the other distinctiveness measures (Tukeys HSD, all pairwise comparisons $p < 0.001$).

Table 1
Proportion of entire tree (PD captured) when 2, 4, 8, and 10 species are selected on a 16-species tree ($n = 5000$ trees)

	Species selected on Yule trees					Species selected on Hey trees				
	2	4	6	8	10	2	4	6	8	10
ES	0.296^{a}	0.505^{a}	0.658^{a}	0.772^{a}	0.859^{a}	0.477^{a}	0.696^{a}	0.811^{a}	0.882^{a}	0.931^{a}
FP	0.283^{c}	0.480^{b}	0.630^{b}	0.748^{b}	0.842^{b}	0.414^{d}	0.610^{d}	0.750^{d}	0.846^{c}	0.912^{b}
PE	0.289^{b}	0.486^{b}	0.633^{b}	0.747^{b}	0.837^{b}	0.444^{b}	0.649^{c}	0.767^{c}	0.846^{c}	0.904^{c}
QE	0.281^{c}	0.482^{b}	0.631^{b}	0.747^{b}	0.840^{b}	0.438^{c}	0.663^{b}	0.785^{b}	0.864^{b}	0.921^{b}
VW	0.280^{c}	0.467^{c}	0.613^{c}	0.727^{c}	0.823^{c}	0.448^{b}	0.663^{b}	0.783^{b}	0.861^{b}	0.916^{b}
AVG	0.270^{d}	0.440^{d}	0.570^{d}	0.679^{d}	0.773^{d}	0.413^{d}	0.591^{e}	0.704^{e}	0.788^{d}	0.855^{d}

For each column, groups that are significantly different at the 0.05 level are designated by different letters.

Across all group sizes and measures, the BIC criterion approach retains only gamma and I_2 as explanatory variables for the amount of PD captured. For example, the model selected for ES at group size 6 was

$$\text{Log(odds)} = 0.215395 + 0.121786 \times \text{gamma} + 3.170172 \times I_2.$$

The amount of PD that was captured by random choice, however, was chiefly dictated by models that contained only gamma and I_c-again the example at group size 6:

$$\text{Log(odds)} = 0.295568 + 0.07805 \times \text{gamma} - 0.451539 \times I_c.$$

The only discrepancies to this pattern are at the largest of the group sizes (i.e. 8 and 10 of 16 tips), where the some models for ES, FP, and PE also contain I_c (Table 2), and at the smallest group size (2 tips chosen), where the model for AVG also contains I_2.

The relative values of the coefficient estimate for gamma were similar for all distinctiveness measures and for random choice, and across all group sizes, showing that this factor of tree shape has little effect on their relative performance. There was greater variation in the coefficient estimates for I_c and I_2, which suggest that they play a key role in determining the relative performance of the different tip-choosing.

Hey trees gave qualitatively similar results with respect to the relative performance of the measures averaged across 5000 trees. ES again captured the most PD but for Hey trees VW and QE captured the next highest amounts, with no significant difference between them, and PE captured the least (see Table 1). FP performed noticeably worse on Hey trees than on Yule trees, capturing only just more than random at the small the group sizes. Because they have very high gamma values the absolute amount of the tree captured by the metrics was higher for Hey trees than for Yule trees, especially for groups of two species, which captured nearly twice as much PD. When selecting large subsets of species, however, the percentages captured were much more similar (Table 1).

The models produced from the Hey data were qualitatively similar in terms of the parameters chosen by the algorithm, differing only slightly in the coefficient values. Given these similarities, we concentrate on the most significant characteristics that are shared by both data sets in our discussion.

Table 2
Parameter estimates of best approximating models for capturing PD by each distinctiveness measure and random choice (AVG), for 8 of 16 tips chosen

	ES	β	SE
ES	u	0.5407	0.0534
	γ	0.1419	0.0111
	I_2	4.3079	0.5633
	I_c	−0.196	0.0175
FP	u	0.4566	0.0587
	γ	0.1089	0.0121
	I_2	0.236	0.5662
	I_c	−0.9051	0.0248
PE	u	0.4344	0.0545
	γ	0.1067	0.0108
	I_2	4.7053	0.5634
QE	u	0.4832	0.0537
	γ	0.1284	0.011
	I_2	3.4899	0.5766
VW	u	0.4345	0.0541
	γ	0.1409	0.0108
	I_2	2.7524	0.5662
AVG	u	0.6023	0.0222
	γ	0.0924	0.0044
	I_c	−0.4515	0.0603

Variables are those included in the top BIC model for each species and significant at $p<0.01$.

β, coefficient; SE, standard error; u, intercept; γ, gamma.

When choosing 25% of the species from each of the 50 "real trees," all distinctiveness measures selected groups with PD higher than the upper 95% c.i. of random choice, on at least 41 of the 50 trees (82%). ES and QE scored equal to, or significantly higher than random, for 92% of the trees, while VW achieved the same on only 86% of the trees. When choosing groups that contained 50% of the species on each tree, ES and PE increased the number of trees on which they capture significantly more than random, with ES increasing from 88% to 90%, and PE from 84% to 90% (Table 3).

Table 3
Percentage of the 50 trees derived from empirical data, on which the measures achieved more then the upper 95% confidence limit of the expected amount of the tree captured (higher), within the confidence limits of the expected amount (same), and lower than the lower confidence limit (lower)

Measure	Lower	Same	Higher
For 25% of the species chosen			
ES	8	4	88
QE	8	8	84
FP	10	6	84
PE	10	6	84
VW	12	4	84
For 50% of the species chosen			
ES	8	2	90
QE	14	2	84
FP	14	2	84
PE	8	2	90
VW	16	2	82

The subset of trees on which none of the measures captured more than random, were both more balanced than trees on which they all captured more than random (ANOVA, $\mathrm{df} = 51$, $p < 0.01$) and had higher gamma (ANOVA, $\mathrm{df} = 48$, $p = 0.05$), in line with the results of the simulation study.

6. Discussion

We highlight four main findings: first, for the small trees considered here, all five distinctiveness metrics perform well at capturing PD for the majority of tree shapes, and generally capture significantly more of the tree than would randomly selecting species (Table 1). The absolute improvement, however, was modest. For example, on the tree of Aves (phylotaxonomy based on Monroe and Sibley, 1993), selecting 50% of the species in the tree using ES captures a further 8 billion years of (concurrent) evolution, which is approximately 9.5% of the total tree when compared to randomly selecting the same number of species (unpublished results). The absolute amount of the tree captured by any of the metrics is a strong function of gamma, as expected. Second, ES consistently captured PD better than all the other measures, and the simplest measure, the PE also performed well. Third, the effects of tree shape on the effectiveness of capturing PD are unsurprising and suggest that distinctiveness may prove a useful metric for conservation. Fourth, while we do not know the extent of "real" tree space, and indeed, how widely our 50 sampled trees represent this space, the concordance of the patterns from our simulations and the "real" trees suggests that our results are likely to be replicated throughout most of the plausible range of "real" tree space.

In contrast to random selection, the metrics are unaffected by the overall balance of tree (I_c). They are affected, however, by the average balance throughout the tree (measured as I_2), and, therefore, for a small proportion of tree shapes they capture PD only as well as random choice. Below, using a graph theoretical approach, we formalize the affect of the balance and tip choice on selecting species at random, and on two of the other measures: PE and ES. Because of its complexity, we defer analytical discussion of QE to a separate full-length treatment; FP has other properties (KH, unpublished results) that also merit a full-length treatment; we do not discuss VW further due to its consistently poor performance and the general problem it has with multiple ties among candidate species.

6.1. Random choice

Random choice does capture a fairly large proportion of total PD (Nee and May, 1997). We note, however, that random choice may not be a realistic model for how species will survive into the future; though the patterns are currently weak, it may be that extinction will be clumped on the phylogenetic tree. In particular, as seen by the negative coefficient estimate (Table 2) random choice performs poorly with increasing tree imbalance. To the extent that published trees are more imbalanced than Yule (Mooers and Heard, 1997; Blum and François, 2006), random choice is compromised.

Consider an ultrametric tree representing n species of which some number, k, can be chosen. All the possible sets of k species with maximal PD can be represented graphically using the following simple method. Locate a time, l, at which k ancestral species exist and partition the n species into the k subsets descendant from these ancestral species. We will refer to these as the optimal k clades: all optimal solutions correspond to selecting one species from each of these clades (Nee and May, 1997).

We first ask: How many of the k clades will be represented by a random selection of species (with the optimal solutions corresponding to all of these clades being represented)? A clade is considered to be represented if at least one species from that clade is conserved.

We denote the probability that a clade of size i is represented by ϕ_i, which corresponds to one minus the probability that all of the k randomly selected species will be chosen from other clades:

$$\phi_i = 1 - \frac{\binom{n-i}{k}}{\binom{n}{k}}, \tag{6}$$

where $\binom{n}{k}$ is the number of ways of selecting k elements from a set of n elements. For brevity, we adopt the usual convention that $\binom{n}{k} = 0$ if $k > n$. We denote the number of the k clades containing i species by a_i. Using this notation the expected number of the k clades that will be represented by a random selection of species is easily expressed:

$$\hat{N} = k - \sum_i \frac{a_i \binom{n-i}{k}}{\binom{n}{k}}. \tag{7}$$

The expected number of clades that are represented depends on the size distribution of the clades (a). The size distribution for which the least number of clades are expected to be represented is that where all clades contain a single species except for one that contains $n{-}k+1$. For this size distribution, Eq. (7) simplifies to

$$\hat{N} = k - \frac{(k-1)\binom{n-1}{k} + \binom{k-1}{k}}{\binom{n}{k}} = 1 + \frac{k(k-1)}{n}. \tag{8}$$

The highest proportion of the k clades is expected to be represented if the species are as evenly distributed amongst the k clades as possible, i.e. if the tree is completely balanced. If n is a multiple of k, Eq. (7) simplifies to

$$\hat{N} = k - \frac{k\binom{n-n/k}{k}}{\binom{n}{k}}. \tag{9}$$

Eqs. (8) and (9) provide a lower and upper bound on the expected proportion of the clades that will be represented by randomly selecting species for any tree. Under the Yule process, the distribution of the sizes of the k clades is geometric (Nee et al., 1992), such that we are nearer the lower than the upper limit (Fig. 2). Note that Hey trees use the same branching patterns as Yule, so any arguments based on tree topology should be applicable to both tree types.

6.2. *Pendant edge*

Using the notation introduced above, when selecting using PE, all of the k clades with size 1 are automatically represented as they have longer PEs than species from any multi-species clades. Hence those clades that are the least likely to be represented by random species selection are guaranteed to be represented under the PE measure. This means PE will do better than random on the most unbalanced tree shapes. However, we must still ask how well the k clades of size >1 are represented under the PE measure. If more than one species is picked from one of these larger clades, some of the other clades must be unrepresented. It is clear that PE will represent more clades if the difference between the longest PEs and the other edges in each clade is as big as possible.

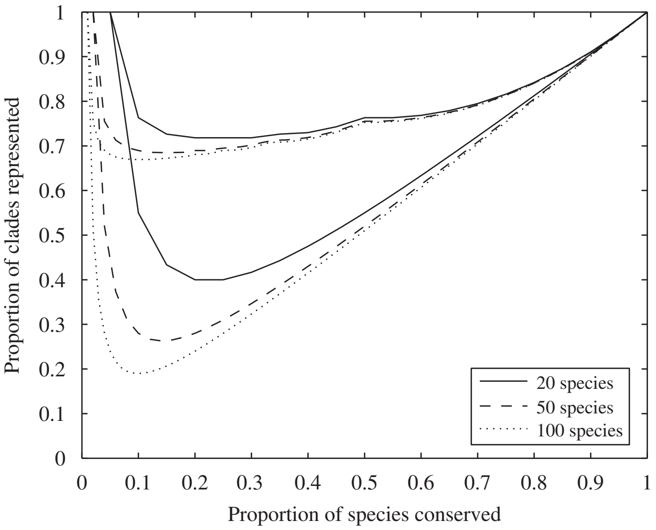

Fig. 2. Expected proportion of the optimal PD captured by random species choice on Yule trees. Under optimal choice, each of k clades (where k is the number of species conserved) should be represented (see text for details). So, for $n = 20$ species, if 5% of species chosen (= 1 species), this must also capture the maximum number of clades, since $k = 1$. Likewise, if all species are chosen, then every one of $k = n$ optimal clades must also be represented. At intermediate values, random species choice will represent $<k$ clades, and this deficit increases with tree size.

To examine this we considered several aspects of edge length probability distributions. These distributions can be obtained by simulating Yule tree data; however, this is a time consuming approach, as large numbers of trees must be simulated to obtain reliable results. Here, we used an alternative analytic scheme for producing these distributions (Hartmann, manuscript, in preparation).

This allows us to make the following observations on Yule trees: first, though the actual topology of the subsets in k affect the distribution of PE among k, larger subsets from k are expected to have a longer PE than smaller subsets from k, and therefore, are more likely to be represented (Fig. 3). In this way, PE will act like random choice. In addition, some of the second-longest edges in some clades may be longer than the longest PEs in other clades, which will lead PE to capture less PD (Fig. 4A). That said, there is good contrast between the longest PEs and the remaining PEs (Fig. 4B), implying that, in general, PE will not choose repeatedly from the same clade.

6.3. *Equal splits*

The ES measure incorporates the PE measure. In fact, the PE generally contributes over half the ES score on Yule trees. On a fully bifurcating tree, if we denote the edges between a PE and the root by λ_0 through to λ_r, the ES measure of a species (Eq. (3)) can be reduced to

$$ES = \sum_i^r \frac{\lambda_i}{2^i}.$$

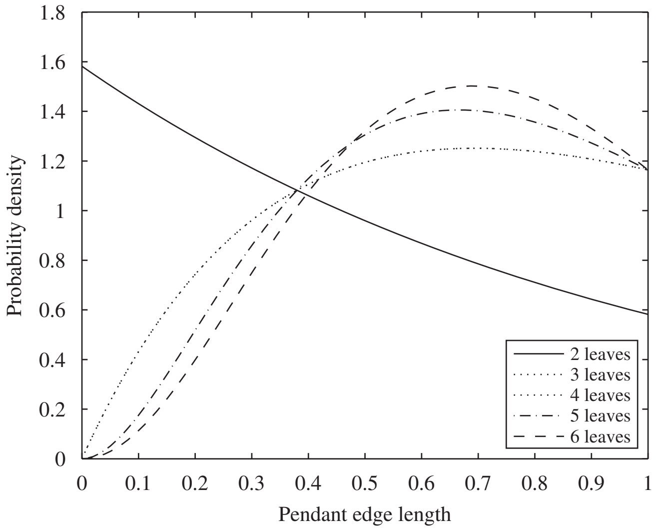

Fig. 3. The probability density of the longest pendant edge for a range of subtree sizes: the density shifts right for larger sub-tree sizes, illustrating that the maximum pendant edge length is more likely in larger sub-clades.

If the branch lengths were equal, this is simply a geometric series; so we obtain

$$ES = 2\lambda\left(1 - \frac{1}{2^{r+1}}\right) \leqslant 2\lambda.$$

The extent to which PEs affect the ES measure depends on the relative magnitudes of PEs and interior edges. If these are of similar magnitude, the PEs will on average contribute half of the ES value. For pure Yule trees, PEs are on average slightly shorter than internal edges as they represent the time from the birth of a species to the present, not the time from the birth of a species to a speciation event.

In any of the k clades, the expected number of internal nodes between a PE and the rest of the tree is lowest for the longest PE due to topological constraints (consider a ladder, or fully imbalanced tree). This increases the contrast between the maximal ES score in each of the k clades and the other ES scores in those clades, when compared to the contrasts found between PE scores in the same situation. This, therefore, directs ES tip choice more efficiently to smaller clades and reduces the number of clades in the tree that are over-represented.

In summary, the ES measure is highly related to the PE measure and for Yule trees is expected to represent a larger proportion of the k clades than the PE measure alone. However, the biases that affect the PE measure are expected to also affect ES, due to their correlation. Findings applicable to ES are also likely to be similarly applicable to FP, due to the fact they are methodologically very similar. It is not known why ES performs better, especially on high gamma trees, but further study is needed to investigate this property.

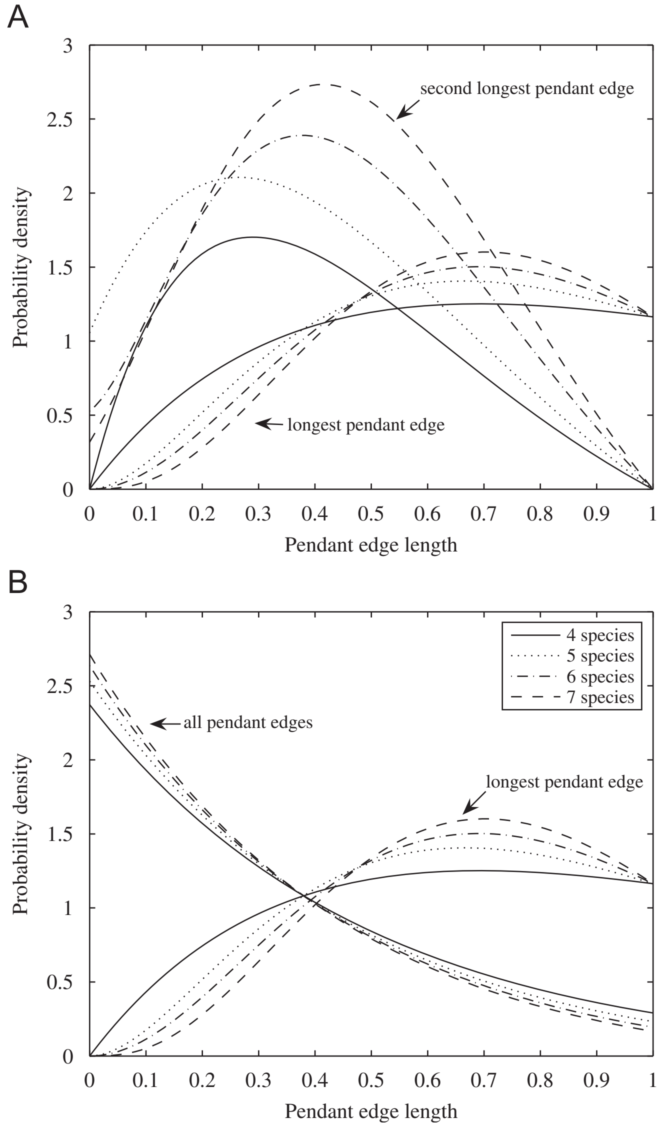

Fig. 4. The ability of pendant edge (PE) measure to capture PD on Yule trees. PE will fail to the extent that it chooses repeatedly from one of the k optimal clades (see text for details). (A) Probability density of the longest (solid lines) and second longest pendant edges (dashed lines) in one of the k clades. The overlap of the two distributions illustrates that it is likely for the sub-tree to be sampled twice by PE. (B) The solid curve is again the probability density for the longest pendant edge, while the dashed line is for all the other pendant edges in a sub-tree. The contrast between the two curves highlights why PE generally over-samples single sub-clades less than random choice.

6.4. *Quadratic entropy*

QE is an explicitly pairwise measure, quite different from the others surveyed here. It is also the most computationally complex of the distinctiveness measures and captures PD at the same rate as PE in the Yule tree data set, but slightly better in the Hey tree data set.

A useful property of QE is that it sets absolute relative relationships between tips (DB, unpublished observations) irrespective of the size of the tree considered: if one species is twice as original as another on a small tree, then if these tips are considered as part of a much larger clade, the relative distinctiveness will remain 2:1. We propose that more work be done investigating the properties of QE, perhaps expanding it from a pairwise to a multi-species framework (cf. Haake et al., 2005; Hartmann and Steel, 2006).

7. Conclusion: implications for conservation

The criterion most often used to prioritize species for conservation is threat status (Possingham et al., 2002). While threatened species tend to come from species-poor groups (Purvis et al., 2000), threat status may not be much more effective at capturing PD than choosing species at random (Redding and Mooers, 2006). Likewise, evolutionary distinctiveness and threat are only very weakly correlated for birds and mammals (AOM and DWR, unpublished observations). If one conservation goal is the preservation of the tree of life, we must attend to explicit tree-based measures.

QE and the Vane-Wright node counting measure are both good at picking out the most relictual species (e.g. they would correctly identify both the tuataras as of highest rank within the squamates) but with the handicap that they do not capture species from across the tree, and therefore, do not capture PD.

PE performed surprisingly well for most of the simulation study and in the 50 sample trees, and using PE as a measure of distinctiveness certainly has many advantages: it is easy to understand, easy to measure, and, perhaps most importantly, it is an absolute measure, meaning any set of species can be compared. Obviously, more work with real trees is required to evaluate how well PE does on average: the relative length of PEs to interior nodes, and the extent to which PE length predicts structure deeper in the tree are open questions (see, e.g. Burlando, 1990). In addition, PE is likely to be very sensitive to alternative species designations (cf. Isaac et al., 2007).

Of the five distinctiveness measures tested here, ES is consistently better than the other measures at capturing PD, at least in the tree space we tested, and is relatively simple to calculate. The cases where we observed FP, VW, PE, or QE to perform better than ES were relatively rare, occurring only in tree shapes where all ranking measures perform badly, i.e. near star-like trees with very low gamma, and where random choice is as good as any other method.

To conclude, we propose that several distinctiveness measures be applied to a wider range of real trees and taxonomies to further explore their properties. We also suggest that the relationship between distinctiveness and other measures of conservation value, particularly conservation status, be explored in more detail. Finally, we must continue the hard work of finding a framework that allows measures of distinctiveness to be compared with other measures of species value, particularly ecological importance, charisma, and costs of recovery and probability of success.

Acknowledgments

We thank Mike Steel and IRMACS for fostering our collaboration, the fab*-lab at SFU for commenting on some of the ideas presented here, and Carolyn Huston for statistical advice. This work was funded by NSERC, Canada, Allan Wilson Centre and the New Zealand Marsden Fund. D.B. and M.D. were supported by Bojar Mahar's Canada Research Chair in Graph Theory at SFU.

Appendix. Supplementary data

Supplementary data associated with this article can be found in the online version at 10.1016/j.jtbi.2007.12.006.

References

Altschul, S.F., Lipman, D.J., 1990. Nature 348, 493–494.

Andelman, S.J., Groves, C., Regan, H.M., 2004. A review of protocols for selecting species at risk in the context of us forest service viability assessments. Acta Oecol. 26, 75–83.

Avise, J.C., 2005. Phylogenetic units and currencies above and below the species level. In: Purvis, A., Gittleman, J.L., Brooks, T. (Eds.), Phylogeny and Conservation. Cambridge University Press, Cambridge, UK, pp. 76–101.

Blum, M.G.B., François, O., 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Syst. Biol. 55, 685–691.

Burlando, B., 1990. The fractal dimension of taxonomic systems. J. Theor. Biol. 146, 99–114.

Colless, D.H., 1982. Phylogenetics: the theory and practice of phylogenetic systematics. Book review. Syst. Zool. 31, 266–276.

Crozier, R.H., 1992. Genetic diversity and the agony of choice. Biol. Conserv. 61, 11–15.

Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61, 1–10.

Faith, D.P., Baker, A.M., 2006. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evol. Bioinform. Online 2, 70–77.

Forest, F., Grenyer, R., Rouget, M., Davies, T.J., Cowling, R.M., Faith, D.P., Balmford, A., Manning, J.C., Proches, S., van der Bank, M., Reeves, G., Hedderson, T.A.J., Savolainen, V., 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. Nature 445, 757–760.

Haake, C.-J., Kashiwada, A., Su, F.E., 2005. The Shapley value of phylogenetic trees. IMW Working Paper No. 363.

Hartmann, K., Steel, M., 2006. Maximizing phylogenetic diversity in biodiversity conservation: greedy solutions to the Noah's Ark problem. Syst. Biol. 55, 644–651.

Hartmann, K., Steel, M., 2007. Phylogenetic diversity: from combinatorics to ecology. Book chapter for: Reconstructing Evolution: New Mathematical and Computational Approaches. In: Gascuel, O., Steel, M. (Eds.). Oxford University Press, Oxford.

Heard, S.B., 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rates. Evolution 50, 141–148.

Hey, J., 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46, 627–640.

Isaac, N.J.B., Turvey, S.T., Collen, B., Waterman, C., Baillie, J.E.M., 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. PLoS ONE 2:e296.

Matsen, F., 2006. A geometric approach to tree shape statistics. Syst. Biol. 55, 652–661.

May, R.M., 1990. Taxonomy as destiny. Nature 347, 129–130.

McPeek, M.A., Brown, J.M., 2007. Clade age and not diversification rate explains species richness among animal taxa. Am. Nat. 169, E97–E106.

Minh, B.Q., Klaere, S., von Haesler, A., 2006. Phylogenetic diversity within seconds. Syst. Biol. 55, 769–773.

Mooers, A.Ø., Heard, S.B., 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72, 31–54.

Moran, P.A.P., 1951. Estimation methods for evolutive processes. J. R. Stat. Soc. B. 13, 141–146.

Monroe Jr., B.L., Sibely, C.G., 1993. A World Checklist Of Birds. Yale University Press, New Haven, London.

Nee, S., May, R.M., 1997. Extinction and the loss of evolutionary history. Science 278, 692–694.

Nee, S., Harvey, P.H., Mooers, A.Ø., 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl Acad. Sci. USA 89, 8322–8326.

Nixon, K.C., Wheeler, Q.D., 1992. Measures of Phylogenetic Diversity. In: Novacek, M.J., Wheeler, Q.D. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 216–234.

Pardi, F., Goldman, N., 2005. Species choice for comparative genomics: no need for cooperation. PLoS Genet. 1, E71.

Pardi, F., Goldman, N., 2007. Resource aware taxon selection for maximising phylogenetic diversity. Syst. Biol. 56 (3), 431–444.

Pauplin, Y., 2000. Direct calculation of a tree length using a distance matrix. J. Mol. Evol. 51, 41–47.

Pavoine, S., Ollier, S., Dufour, A.B., 2005. Is the originality of a species measurable? Ecol. Lett. 8, 579–586.

Peng, R.D., 2005. Simpleboot: simple bootstrap routines (online). Available from: ⟨http://Sandybox.Typepad.Com/Software/⟩ (last accessed 18 July 2006).

Possingham, H.P., Andelman, S.J., Burgman, M.A., Medellin, R.A., Master, L.L., Keith, D.A., 2002. Limits to the use of threatened species lists. Trends Ecol. Evol. 17, 503–507.

Purvis, A., Agapow, P.M., Gittleman, J.L., Mace, G.M., 2000. Non-random extinction and the loss of evolutionary history. Science 288, 328–330.

Pybus, O.G., Harvey, P.H., 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. London Biol. B 267, 2267–2272.

Rao, C.P., 1982. Diversity and similarity coefficients: a unified approach. Theor. Pop. Biol. 21, 24–43.

Redding, D.W., 2003. Incorporating genetic distinctness and reserve occupancy into a conservation priorisation approach. Masters Thesis, University Of East Anglia, Norwich, UK.

Redding, D.W., Mooers, A.O., 2006. Incorporating evolutionary measures into conservation prioritization. Conserv. Biol. 20, 1970–1978.

Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, New York.

Steel, M., 2005. Phylogenetic diversity and the greedy algorithm. Syst. Biol. 54, 527–529.

Steel, M., Mimoto, A., Mooers, A.O., 2007. Hedging one's bets: quantifying a taxon's expected contribution to future phylogenetic diversity. Evol. Bioinform. 3, 237–244.

Vane-Wright, R.I., Humphries, C.J., Williams, P.H., 1991. What to protect—systematics and the agony of choice. Biol. Conserv. 55, 235–254.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York.

Vos, R., 2006. Phylogenetic inference: the 'big tree problem'. Ph.D. Thesis, Simon Fraser University, Burnaby, Canada.

Weitzman, M.L., 1998. The Noah's Ark problem. Econometrica 66, 1279.

Yule, G.U., 1924. Philos. Trans. R. Soc. London A 213, 21–87.