The most “original” species often capture more phylogenetic diversity than expected

David W. Redding\textsuperscript{1,2,7}, Klaas Hartmann\textsuperscript{3,7}, Aki Mimoto\textsuperscript{2,7}, Drago Boka\textsuperscript{1,4}, Matt DeVos\textsuperscript{2,5} and Arne Ø. Mooers\textsuperscript{1,2,6}

\textsuperscript{1}BISC and \textsuperscript{2}IRMACS and \textsuperscript{5}Mathematics, Simon Fraser University Burnaby, Canada
\textsuperscript{3}Biomathematics Research Center, University of Canterbury, Christchurch, New Zealand
\textsuperscript{4}Institute of Mathematics, Physics, and Mechanics, Ljubljana, Slovenia
\textsuperscript{6}Institute for Advanced Study, Berlin, Germany
\textsuperscript{7}These authors contributed equally

Word Count: 5825

Email: dredding@sfu.ca; amooers@sfu.ca

Keywords: conservation, distinctiveness, evolutionary worth, genetic uniqueness, originality measures, phylogenetic diversity.
Abstract

Originality is a measure of how evolutionarily isolated a species is relative to other members of its clade. Recently, originality 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. So, on the tree of all the world’s birds, a pairwise measure based on quadratic entropy captured significantly less PD than did random choice, while a simple tree apportioning measure captured significantly more. We used simulation and simple graph theory to explore the properties of these and other measures with reference to phylogenetic tree shape. Overall, the originality measures do better 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. Taken together, this suggests that originality 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 and may provide a simple transparent alternative to more complicated conservation prioritization approaches.

Introduction

With increasing extinction there is a pressing need to effectively prioritize species for conservation. Many nonexclusive currencies are used, e.g. threat status, ecological importance, social or intrinsic value, financial cost (for discussion, see Crozier et al 1997; 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 phylogenetically distinct species in a tree and prioritizing sets of species that best represent the whole tree.

Phylogenetic distinctness has been approached from two different angles. The idea of comparing the relative phylogenetic diversity 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 phylogenetic diversity (PD) of a set of species is the sum of the branch lengths of the tree containing those species and the root (see Figure 1 and Faith and Baker, 2006).

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 & Wheeler 1992, Pavoine et al. 2005, Redding 2003, Redding & Mooers 2006, Isaac et al., manuscript). Both approaches use information about the relatedness among tips (usually species) on a tree, but whereas phylogenetic diversity is a group measure, the latter (‘originality’ sensu Pavoine et al., 2005) is a species-specific property.

We can illustrate this by considering 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 a species from each order were equally threatened, priority should go to a tuatara before any lizard or snake species, because the tuatara species are highly distinct and contain a disproportionately large proportion of the phylogenetic diversity contained within the two groups. If only two species of the ~6201 species were to be preserved (an unlikely scenario), the max PD set would include one of the two tuataras, and one Squamate.
Although useful for ranking predefined areas for conservation, 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. Alternative variations of PD that overcome this have been proposed (Hartmann et. al., manuscript) but these are relatively complex.

The initial attempts to attribute a score of evolutionary distinctness to individual species (Vane-Wright 1991, May 1990, Nixon and Wheeler 1992) used only tree topology, and relied on the fact that basal and evolutionary isolated species have fewer nodes between the tip and the root. This type of approach suffers from poor resolution among species (i.e. many tied values) and also loses any information conferred by branch lengths (Pavoine et al. 2005). Recent papers (Pavoine et al. 2005, Redding and Mooers 2006; Weitzman 1998) have suggested originality measures (outlined below) that use both topology and internal branch lengths to establish how isolated species are. All these 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.

A common, default approach to prioritizing species for conservation is by threat status (Possingham et al. 2002). Despite threatened species tending to come from species poor groups (Purvis et al. 2000) this may not be much more effective at capturing PD than choosing species at random (Redding and Mooers 2006).

Using simple algorithms (Steel 2005) sets of species with maximal PD can be found. This can be extended to include species survival probabilities and conservation costs and budgets (Weitzman 1998, Hartmann and Steel, 2006, Pardi and Goldman 2005). Many current conservation approaches (e.g. endangered species lists) rely, however, on having
species ranked in order of priority. Current optimal approaches offer no such order and may in any case be too complex for practical use in conservation organizations.

To overcome this, species within any optimal set chosen can be ordered by arranging them according to their originality, or, alternatively, species can 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 these approaches further complicate matters when implementing such systems 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 prioritise and other complex factors such as cost of conserving individual species are considered. (Hartmann and Steel, 2006).

The recently developed species-specific measures of evolutionary distinctness may, in comparison, represent a flexible and transparent conservation tool to promote “evolutionary value” in the current legislative climate. Importantly, though, if using these measures to choose the set of most original species also meant choosing a phylogenetically diverse *set* of species, such species-specific measures would be doubly useful, highlighting the most individually distinct species and helping preserve more of the tree of life.

Tree shape is likely to be an important factor in determining how effectively originality measures capture the tree. We outline three measures of tree shape and how we think they will affect PD capture.
The balance of the tree towards the root (Ic; 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 originality scores (e.g. there will be the same number of nodes between the tips and the root). Originality 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 $n=16$ tip Yule trees, this relationship is very strong: Pearson’s Rho = 0.988, $n = 5005$, $p<0.001$). Cherries will have the same originality 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. All the species in the species-poor side of the tree will, most likely, have lowest number of nodes between tip and root. A simple originality measure such as VW (Vane Wright et al, 1991) will most likely exhaustively choose species from just the species-poor side of tree without crossing the root.

The final aspect of a tree branching structure that is likely to affect capture rate is average node depth, 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 $<<0$), where species contribute equally to PD value ($\sim$PD/n): each species contributes the same amount to total PD; at the other extreme, where most of the nodes are at the present (gamma $>>0$), the
first species chosen captures ~PD/2, and most additional species contribute little. In both cases these 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 originality metrics captures PD, and to explore how tree shape might affect this efficiency, we undertook simple simulation and then formalized some of the results using a graph-theoretical approach.

The measures

There exists a small set of older, node-based measures of originality (reviewed in Pavoine et al., 2005). 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. Given a tree $T$ and a species $i$, let $p(T,i,r)$ denote a path of edges from root to the species, and $\text{deg}_{\text{out}}(e)$ be the number of lineages emanating from node $e$ (two for a bifurcating tree):

$$VW(T,i) = \frac{1}{\sum_{e \in p(T,i,r)} \text{deg}_{\text{out}}(e)}$$  

(1)
Quadratic Entropy (QE; Rao 1982) was initially proposed as a measure of the biodiversity for a species assemblage. Rao’s QE metric 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 originality measure: if species-pairs are sampled in proportion to their originality, pairwise phylogenetic distance is maximized. So, Pavoine et al. (2005) assign weights \( p_i \) to maximize \( Q(p) \):

\[
Q(T, p) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \sum_{k \in p(T, i, j)} \lambda_k
\]

(2)

Given a tree \( T \) and a probability distribution, \( p \) is the solution that maximizes the QE value. Each member of \( p \) is associated with a single species. We draw attention to the fact that QE is a pair-wise measure. We (DB and MD) implemented a new algorithm for QE in C++ that runs in linear time (details to be supplied elsewhere; program available on request.)

Equal-splits (ES), on the other hand, simply distributes the phylogenetic 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 equal-splits value from every internal branch between a species tip and the root is a representation of the unique amount of evolutionary 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 equal-splits measure across the tips equals the total phylogenetic diversity of the tree (Pauplin 2000; Semple & Steel 2004):
\[ ES(T, i) = \sum_{e \in p(T, i, r)} \left( \lambda_e \prod_{f \in p(T, i, e)} \frac{1}{(\text{deg}-1)_f} \right) \]  

(3)

where for a tree \( T \) and a species \( i \), \( p(T, i, r) \) denotes the path of edges from root to species \( i \), and “deg” is the number of degrees of the vertex at the outward end of edge \( f \). Equal splits for species \( i \) is the sum of all edge lengths \( \lambda_e \) divided by the product of all the outward degrees (ie, for a bifurcating tree, deg-1=2) for vertices along the path from root to species.

\[ ES \] is closely related (Pearson correlation coefficient = 0.96, \( n=5000 \), \( p<0.001 \)) to another measure called Evolutionary Distinctness (ED, originally SSPD in Redding 2003), which has been used to measure the originality of species in a recently suggested prioritization framework (Isaac et al, manuscript). Due to its similar behavior to ES for most tree shapes, and given its ability to capture PD is very slightly lower than ES, it was not considered further in this study.

Unlike the first three measures, the age of the species, or its Pendant Edge value (PE, 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. 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 redated.

\[ PE(T, i) = PD(T) - PD(T - i) \]  

(4)

Given a tree \( T \) and a species \( i \), the PE value for the species \( i \) is given by its individual contribution to PD (for example the PE value for species A in Figure 1 is 1 million years).

The expected value if species were chosen randomly was used here as a benchmark (AVG, following Nee and May, 1997, but calculated following Appendix 1)
Methods

Our primary dataset 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 coalescent (Moran, 1958) and have Yule topologies, but different waiting times, with more splits occurring near the present.

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 ultrametric trees.

For each simulated tree, we used the four metrics to rank the 16 species (ES, 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 was random; see Appendix One).

We first asked if we could predict the amount of PD captured (using the four Originality measures and the AVG algorithm) using a series of exploratory logistic regression models, with the y variable being percentage of the tree captured, using a logit link and a binomial distribution, and our three estimates of tree shape: Ic (Colless 1982; Heard, 1996); I2 (Heard and Mooers 1997; Matsen, 2006); and Gamma (Pybus and Harvey, 2000). In total five separate models were constructed using 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 6 measures (the 4 originality metrics, the average (expected) proportion, and for comparison, the optimal proportion possible) for each
of the 5000 trees, yielding n=833 in our models. We concentrate on the differences between random choice and the choice using the originality metrics.

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

Finally, we applied the four originality metrics to a “real tree” representing relationships among birds, the Sibley and Ahlquist phylotaxonomy (Monroe and Sibley, 1993, n=9702 spp) and asked how much PD was represented by the most original set of species at different selection sizes. In order to help visualize how the rank orders given by the different originality measures translated into the different PD capture rates, we plotted the unique ID number of a bird species, assigned sequentially to a simple taxonomic list - such that species that are close together on the tree were given similar values - against the rank that the species would be chosen by each of the metrics.

\textbf{Results}

For small trees, all four originality 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 originality 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 four measures doing better than random in this regard; corresponding numbers for 8 species are 72-77%; (Table 1). The Equal-Splits measure captured more PD than the other originality measures (Tukey’s HSD, all pairwise comparisons p<0.001).

Across all group sizes the BIC criterion approach choose very similar models for the originality measures; they contained gamma and I2. For example the model selected for ES at group size 6 was

\[
\log (\text{odds}) = 0.215395 + 0.121786 \times \text{gamma} + 3.170172 \times I2
\]

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

\[
\log (\text{odds}) = 0.295568 + 0.07805 \times \text{gamma} - 0.451539 \times Ic
\]

The only discrepancies to this pattern are at the largest of the group sizes we modeled (i.e. 8 and 10 of 16 tips) the models for each of the ES and PE also contain Ic and at the smallest group size (2 tips chosen) the model for AVG also contains I2 (Table 2).

The relative values of the coefficient estimates are similar for gamma, showing that this factor of tree shape has little effect on the relative performance of the measures and random choice. Ic and I2 alternatively had greater variation in coefficients, which suggests they play a key role in determining the relative performance of the different tip choosing methods on different shaped trees.

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). 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 2 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 coefficients values. Given
these similarities, we concentrate on the most significant characteristics that are shared by
both datasets in our discussion.

On the large Sibley and Ahlquist bird tree, the measures performed differently from
one another (Figure 2). While Equal Splits (Redding and Mooers, 2006) and Pendant Edge
(Altschul and Lipman, 1990) produced rankings that captured more PD than a random
selection of species, Quadratic Entropy (Pavoine et al., 2005) and node counting (Vane-
Wright et al., 1991) captured significantly less. Figure 3 shows why: QE and VW choose
repeatedly from a small subsection of the tree, while ES and to a lesser extent PE choose
from a broader taxonomic range.

Discussion

We highlight three main findings: First, for the small trees considered here, all four
originality 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. And, on the tree of Aves, selecting
50% of the species in the tree using ES captures a further 8 billion years of (concurrent)
evolution when compared to randomly selecting the same number of species. 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. Third, the effects of tree shape on the effectiveness of capturing PD are unsurprising and suggest that originality may prove a useful metric for conservation.

In contrast to random selection, the metrics are unaffected by the overall balance of tree (Ic). They are affected, however, by the average balance throughout the tree (measured as I2) 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 Random selections, and on the Pendant Edge and Equal Splits measures. Because of its complexity, we defer analytical discussion of QE to a separate full-length treatment and we do not discuss VW further due to its consistently poor performance and lack of differentiation among candidate species.

**Random Choice**

Random choice does capture a fairly large proportion of total PD (Nee and May, 1997). We note that random choice may not be a realistic model for how species will be conserved into the future; though the patterns are currently weak, it may be that extinction will be clumped on the phylogenetic tree. Furthermore, as seen by the negative coefficient estimate (Table 2) random choice performs poorly with increasing tree imbalance. 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 (see Figure 1). We will
refer to these as the optimal $k$ clades as all optimal solutions correspond to selecting one species from each of these clades.

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 $\phi_i$, which corresponds to one minus the probability that all of the $k$ randomly selected species will be chosen from other clades:

$$\phi_i = 1 - \binom{n-i}{k} \binom{n}{k}$$

(5)

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:

$$\hat{N} = k \sum_i a_i \binom{n-i}{k} \binom{n}{k}$$

(6)

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 equation 6 simplifies to:

\[
\hat{N} = k - \frac{(k-1) \binom{n-1}{k} + \binom{k-1}{k}}{n} \\
= 1 + \frac{k(k-1)}{n}
\] (7)

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 \( k \) equation 6 simplifies to:

\[
\hat{N} = k \frac{\binom{n-n/k}{k}}{n} 
\] (8)

Equations 7 and 8 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, meaning random choice will do quite poorly (Figure 4). 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. To the extent that published trees are more imbalanced than Yule (Mooers and Heard, 1997; Blum et al., 2006), random choice is further compromised.

**Pendant Edge**
When selecting using PE, all of the $k$ clades with size 1 are automatically represented as they have longer pendant edges 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 clades from the set \{1.. $k$\} 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 pendant edges and the other edges in each clade is as big as possible.

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 time a consuming approach as large numbers trees must be simulated to obtain reliable results. Here, we used an alternative analytic scheme for producing these distributions (Hartmann, manuscript, in prep.).

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 longest pendant edge than smaller subsets from $k$ and therefore are more likely to be represented (Figure 5). 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 pendant edges in other clades, which will lead PE to capture less PD (Figure 6A). That said, there is good contrast between the longest pendant edges and the remaining pendant edges (Figure 6B), implying that, in general, PE will not chose repeatedly from the same clade.
**Equal Splits**

The equal-splits (ES) measure incorporates the PE measure. In fact, the pendant edge generally contributes over half the ES score on Yule trees. On a fully bifurcating tree, if we denote the edges between a pendant edge and the root by $\lambda_0$ through to $\lambda_r$, the ES measure of a species (Equation 3) can be reduced to:

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

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

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

The extent to which pendant edges effect the equal splits measure depends on the relative magnitudes of pendant edges and interior edges. If these are of similar magnitude the pendant edges will on average contribute half of the $ES$ value. For Yule trees, pendant edges 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 pendant edge and the rest of the tree is lowest for the longest pendant edge due to topological constraints (consider a ladder, or fully imbalanced, tree). This further increases the contrast between the maximal ES score in each of the $k$ clades and the other ES scores 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 equal splits measure is highly related to the pendant edge measure and for Yule trees is expected to represent a larger proportion of the $k$ clades than the pendant
edge measure alone. However, the biases that affect the PE measure are expected to also affect ES, due to their correlation.

**Quadratic Entropy**

QE is an explicitly pair-wise measure, quite different from the others surveyed here. It is also the most computationally complex of the originality measures and captures PD at the same rate as PE in the Yule tree dataset, but slightly better in the Hey tree dataset. On the bird tree QE behaves most like VW as it repeatedly samples from the Oaves, the most basal group, poorly capturing PD. All measures are caught out by cherries, but QE and VW are also caught out by more complicated ‘snags’ on larger trees.

A useful property of QE is that it sets absolute relative relationships between tips (DB and MdV, unpublished results) 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 originality will remain 2:1. We propose that more work be done investigating the properties of QE, perhaps expanding it from a pairwise to a multispecies framework (c.f. Haake et al., 2005; Hartmann and Steel, 2006).

**Conclusion: connotations for conservation**

Quadratic entropy and the Vane-Wright node counting measure are both good at picking out the most relictual species (e.g. they would correctly identify the two 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 its length.
PE performed surprisingly well for most of the simulation study, and using PE as a measure of originality 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 because the relative length of pendant edges to interior nodes, and the extent to which pendant edge length predicts structure deeper in the tree is an open question (see, e.g., Burlando 1990).

Of the four originality measures tested here, Equal Splits (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 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.

The originality scores we surveyed do very well on both the large tree of the Aves, and on the simulated trees. Our data show that trees with medium gamma and moderate to high imbalance will benefit the most from using the originality measures. Interestingly, a large majority (approximately XX%) of trees measured by Blum et al. (2006) fall within this tree space for topologies, suggesting that originality measures have high potential to increase the capture of PD for trees based on empirical data.

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

Acknowledgements

We thank Mike Steel and IRMACS for fostering our collaboration, the fab*-lab at SFU for commenting on some of the ideas presented here, Carolyn Huston for statistical advice. This work was funded by NSERC Canada, the New Zealand Marsden Fund, and XXX (Drago, Matt).

Literature Cited


Pardi F & Goldman N. 2005. Species Choice For Comparative Genomics: No Need For Cooperation. Plos Genetics 1:E71


Figure 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.
Figure 2. The amount of PD captured by the three different originality measures when ranking the 9702 bird species in the Monroe and Sibely (1990) phylotaxonomy.
Figure 3. A representation of the pattern on the Sibley and Ahlquist phylotaxonomy when species are chosen based on the Originality metrics. VW and QE continually sample from one portion of the tree, leading to autocorrelation (and reduced effectiveness, see Figure 2), while ES and PE sample more widely. Random sampling would produce a pattern of white noise.
Figure 4. 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 clades, since \( k =1 \). Likewise, if all species are chosen, then every one of the \( k=n \) optimal clades must also be represented. At intermediate values, random species choice will represent \(<k \) clades, and this deficit increases with tree size.
Figure 5. The probability density of the longest pendant edge for a range of subtree sizes: the density shifts right for larger subtree sizes, illustrating that the maximum pendant edge length is more likely in larger subclades.
Figure 6. The ability of the Pendant Edge 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 subtree 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 subtree. The contrast between the two curves highlights why PE generally oversamples single subclades less than random choice.
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). For each column, groups that are significantly different at the 0.05 level are designated by different letters.

<table>
<thead>
<tr>
<th></th>
<th>Species selected on Yule Trees</th>
<th>Species selected on Hey Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>ES</td>
<td>0.296&lt;sup&gt;b&lt;/sup&gt; 0.505&lt;sup&gt;b&lt;/sup&gt; 0.658&lt;sup&gt;b&lt;/sup&gt; 0.772&lt;sup&gt;b&lt;/sup&gt; 0.859&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.477&lt;sup&gt;b&lt;/sup&gt; 0.696&lt;sup&gt;b&lt;/sup&gt; 0.811&lt;sup&gt;b&lt;/sup&gt; 0.882&lt;sup&gt;b&lt;/sup&gt; 0.931&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>PE</td>
<td>0.289&lt;sup&gt;c&lt;/sup&gt; 0.486&lt;sup&gt;c&lt;/sup&gt; 0.633&lt;sup&gt;c&lt;/sup&gt; 0.747&lt;sup&gt;c&lt;/sup&gt; 0.837&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.444&lt;sup&gt;c&lt;/sup&gt; 0.649&lt;sup&gt;d&lt;/sup&gt; 0.767&lt;sup&gt;d&lt;/sup&gt; 0.846&lt;sup&gt;d&lt;/sup&gt; 0.904&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>QE</td>
<td>0.281&lt;sup&gt;d&lt;/sup&gt; 0.482&lt;sup&gt;e&lt;/sup&gt; 0.631&lt;sup&gt;c&lt;/sup&gt; 0.747&lt;sup&gt;c&lt;/sup&gt; 0.840&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.438&lt;sup&gt;d&lt;/sup&gt; 0.663&lt;sup&gt;c&lt;/sup&gt; 0.785&lt;sup&gt;c&lt;/sup&gt; 0.864&lt;sup&gt;c&lt;/sup&gt; 0.921&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>VW</td>
<td>0.280&lt;sup&gt;d&lt;/sup&gt; 0.467&lt;sup&gt;d&lt;/sup&gt; 0.613&lt;sup&gt;d&lt;/sup&gt; 0.727&lt;sup&gt;d&lt;/sup&gt; 0.823&lt;sup&gt;d&lt;/sup&gt;</td>
<td>0.448&lt;sup&gt;c&lt;/sup&gt; 0.663&lt;sup&gt;c&lt;/sup&gt; 0.783&lt;sup&gt;c&lt;/sup&gt; 0.861&lt;sup&gt;c&lt;/sup&gt; 0.916&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>AVG</td>
<td>0.270&lt;sup&gt;e&lt;/sup&gt; 0.440&lt;sup&gt;e&lt;/sup&gt; 0.570&lt;sup&gt;e&lt;/sup&gt; 0.679&lt;sup&gt;e&lt;/sup&gt; 0.773&lt;sup&gt;e&lt;/sup&gt;</td>
<td>0.413&lt;sup&gt;c&lt;/sup&gt; 0.591&lt;sup&gt;e&lt;/sup&gt; 0.704&lt;sup*e&lt;/sup&gt; 0.788&lt;sup&gt;e&lt;/sup&gt; 0.855&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Table 2. Parameter estimates of best approximating models for capturing PD by each originality measure and random choice (AVG), organized by number of tips chosen (2-10 out of 16). Variables are those included in the top BIC model for each species and significant at P < 0.01.

<table>
<thead>
<tr>
<th></th>
<th>ES</th>
<th>PE</th>
<th>QE</th>
<th>VW</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>-0.6149</td>
<td>0.0163</td>
<td>u</td>
<td>-0.6745</td>
<td>0.0366</td>
</tr>
<tr>
<td>γ</td>
<td>0.1166</td>
<td>0.0021</td>
<td>γ</td>
<td>0.1166</td>
<td>0.0021</td>
</tr>
<tr>
<td>I₂</td>
<td>1.3141</td>
<td>0.1774</td>
<td>I₂</td>
<td>1.3141</td>
<td>0.1774</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>-0.1736</td>
<td>0.0505</td>
<td>u</td>
<td>-0.1477</td>
<td>0.0518</td>
</tr>
<tr>
<td>γ</td>
<td>0.1146</td>
<td>0.0165</td>
<td>γ</td>
<td>0.0948</td>
<td>0.0162</td>
</tr>
<tr>
<td>I₂</td>
<td>2.2746</td>
<td>0.5264</td>
<td>I₂</td>
<td>1.3007</td>
<td>0.5315</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>0.2154</td>
<td>0.0511</td>
<td>u</td>
<td>0.1717</td>
<td>0.0523</td>
</tr>
<tr>
<td>γ</td>
<td>0.1218</td>
<td>0.0167</td>
<td>γ</td>
<td>0.0983</td>
<td>0.0165</td>
</tr>
<tr>
<td>I₂</td>
<td>3.1702</td>
<td>0.5347</td>
<td>I₂</td>
<td>2.8112</td>
<td>0.5375</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>0.5407</td>
<td>0.0534</td>
<td>u</td>
<td>0.4344</td>
<td>0.0545</td>
</tr>
<tr>
<td>γ</td>
<td>0.1419</td>
<td>0.0111</td>
<td>γ</td>
<td>0.1067</td>
<td>0.0108</td>
</tr>
<tr>
<td>I₂</td>
<td>4.3079</td>
<td>0.5633</td>
<td>I₂</td>
<td>4.7053</td>
<td>0.5634</td>
</tr>
<tr>
<td>Ic</td>
<td>-0.1960</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
<td>β</td>
<td>SE</td>
</tr>
<tr>
<td>(10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u</td>
<td>0.8301</td>
<td>0.0578</td>
<td>u</td>
<td>0.7681</td>
<td>0.0588</td>
</tr>
<tr>
<td>γ</td>
<td>0.1346</td>
<td>0.0121</td>
<td>γ</td>
<td>0.1044</td>
<td>0.0116</td>
</tr>
<tr>
<td>I₂</td>
<td>5.4141</td>
<td>0.6808</td>
<td>I₂</td>
<td>4.6231</td>
<td>0.6755</td>
</tr>
<tr>
<td>Ic</td>
<td>-0.3743</td>
<td>0.1580</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

β, coefficient; SE, standard error; u, intercept; γ, gamma
APPENDIX 1 Calculating E(PD) from random samples of Species

The expected PD of a subtree of size \( n \) (denoted AVG in the main text) is:

\[
PD(T, n) = \sum_{e \in T} \left( 1 - \frac{C(|L_e'|, n)}{C(|L|, n)} \right) \lambda_e \tag{A1.1}
\]

Where for each edge \( e \) in tree \( T \), the associated edge length \( \lambda_e \). \( C(|L|, n) \) is the number of combinations of size \( n \) possible for the number of leaves present \( |L| \), and \( C(|L_e'|, n) \) is the number of combinations of size \( n \) possible for the number of leaves not subtended by \( e \).

Given a rooted tree \( T \), the tree will have a set of leaves \( L \) and set of edges \( E \), where each edge \( e \), has a length \( \lambda_e \). The PD of such a tree is equal to its length (Faith 1992):

\[
PD(T) = \sum_{e \in T} \lambda_e \tag{A1.2}
\]

A particular edge \( e \) is included in the full sum of PDs for every subtree that contains that edge. Thus, the expected PD can be calculated by considering each edge in \( T \) and its proportional presence \( p_e \), in the full set of subtrees.

\[
\overline{PD}(T, n) = \sum_{e \in T} p_e \lambda_e \tag{A1.3}
\]

An edge \( e \) will be present in all subtrees that contain a leaf subtended by \( e \). Let \( L_e'' \) be the set of leaves that are not subtended by \( e \). Then, the number of subtrees that do not use \( e \) is given by counting subtrees of size \( n \) chosen solely from leaves that are not subtended by \( e \) given by \( C(|L_e''|, n) \). Then, the number of subtrees of size \( n \) that contain \( e \), \( N(T, n, e) \) is given by:
\[ N(T, n, e) = C(L_e, n) - C(L_n, n) \]  \hspace{1cm} \text{A1.4}

The expected PD can now be calculated directly:

\[ \overline{PD}(T, n) = \sum_{e \in T} \left( 1 - \frac{C(L_e, n)}{C(L_n, n)} \right) \lambda_e \]  \hspace{1cm} \text{A1.5}

When more than one species share the same originality score and are of identical rank it is not possible to choose a particular sequence of species for calculation of PD. In this situation we calculate the average PD conserved by randomly choosing among tied members repeatedly.