6

PHYLOGENETIC DIVERSITY: FROM COMBINATORICS TO ECOLOGY

Klaas Hartmann and Mike Steel

Abstract

The phylogenetic diversity (PD) of a set of taxa contained within a phylogenetic tree is a measure of the biodiversity of that set. PD has been widely used for prioritizing taxa for conservation and is the basis of the 'Noah's Ark Problem' in biodiversity management. In this chapter we describe some new and recent algorithmic, mathematical, and stochastic results concerning PD. Our results highlight the importance of considering time scales and survival probabilities when making conservation decisions. The loss of PD under a simple extinction process is also described for any given tree—this provides contrasting results depending on whether extinction is measured as function of time or of the number of lost species. Lastly we explore a very different application of PD, its use for reconstructing trees and the associated mathematical properties. The wide range of applications in this chapter shows the usefulness of PD for exploring phylogenetic tree structure with further applications sure to follow.

6.1 Introduction and terminology

Phylogenetic diversity (PD) is a measure of the evolutionary history spanned by a set of taxa within a larger phylogenetic tree. Briefly, the PD score of a subset of taxa, S, is the sum of the lengths of those edges of the tree that span S (a more precise definition follows later). It has been used as a comparative measure in biodiversity conservation, following its introduction by Dan Faith in 1992 [14]. Subsequent authors (see, for example [2, 10, 29, 36, 52] and the references therein) have further explored its application to biodiversity assessment and conservation. Properties of phylogenetic diversity have also been applied recently by Pardi and Goldman [34] to shed light on the relative merits of cooperative versus greedy strategies for taxa sampling for genomic sequencing.

In this chapter we explore some mathematical and stochastic properties of phylogenetic diversity in three different settings: biodiversity conservation, patterns in biodiversity loss, and its relevance for tree reconstruction.

First we explain how PD satisfies two combinatorial properties, including a certain greedoid-type inequality (from [46]) which is algorithmically useful for

selecting sets of taxa to optimize PD. We also consider some taxon-specific indices based on PD that give an indication of the relative distinctiveness of each taxon in a tree. We show that these indices have some shortcomings when used to guide biodiversity conservation and consider a framework that overcomes some of these limitations. This framework is the 'Noah's Ark Problem' (NAP) introduced by Weitzman [52]. In the NAP each taxon has survival probabilities and conservation costs. The NAP seeks the optimal allocation of limited funds to conserving taxa such that the expected remaining future PD is maximized. Under certain restrictions a fast, 'greedy' algorithm provides a solution to this problem. We extend one such result from [46] and use this extension to investigate the management time scale that is implicit in the NAP.

We then investigate the loss of PD as taxa randomly become extinct. Nee and May [30] investigated this process for randomly generated trees. They found a characteristic concave shape in the relationship between expected PD and the proportion of taxa deleted. We describe a result that shows how this is the expected behaviour for any given tree (with any given branch lengths). This indicates that most of the loss of PD comes near the end of an extinction process. However, if one examines the behaviour of expected PD as a function of time, then a contrasting (partially convex, rather than concave) relationship emerges.

Finally we examine the role of PD in the reconstruction of phylogenetic trees. PD estimates for triples (or larger numbers) of taxa have recently been investigated as a way to refine the popular Neighbor-Joining algorithm; it is also possible to consider PD over any abelian group. We describe the mathematical properties of PD in these two settings.

6.2 Definitions and combinatorial properties

Mostly we follow the notation used by Semple and Steel [42]. We let $\mathcal{T}$ denote a *phylogenetic X–tree*, that is, a tree whose leaves comprise the set X of taxa (generally species or populations) under study, and whose remaining vertices (nodes) are of degree at least 3 (the degree of a vertex is the number of edges that are incident with it). The vertices at the tips are called *leaves*. If all the non-leaf vertices in a tree have three incident edges the tree is said to be *fully resolved* (sometimes called 'binary'—these are the trees without polytomies, and so are maximally informative). We also deal with *rooted trees* which have some vertex (often the mid-point of an edge) distinguished as a root vertex. If we direct all the edges of the tree away from the root (i.e. so they are consistent with a time direction if the root is the ancestral taxon) then we can talk about the *clusters* of the tree—the subsets of X that lie below the different vertices of the tree. It is a classical result that any rooted phylogenetic tree can be uniquely reconstructed from its set of clusters. Often the edges of the trees (rooted or unrooted) will have a *(branch) length*—corresponding perhaps to the expected amount of evolutionary change on that edge.

Given a (rooted or unrooted) phylogenetic X–tree, $\mathcal{T}$, with branch lengths, and given a subset Y of X, the *phylogenetic diversity* (PD) of Y, denoted $PD(Y)$

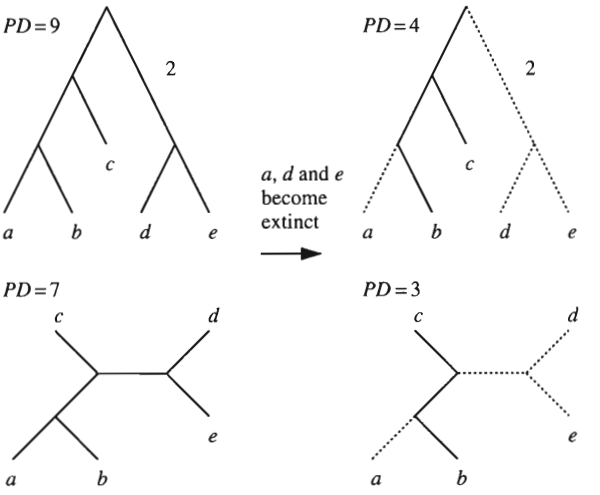


FIG. 6.1. The PD of the trees on the left is calculated by summing the edge lengths. All edges are length 1 except for the long edge on the rooted tree which has length 2. The trees on the right show which edges are considered to remain (solid lines) after taxa a, d, and e become extinct. The PD of these trees is the sum of the remaining edges.

is the sum of the lengths of all the branches that connect the leaves in Y (and also the root of $\mathcal{T}$ if $\mathcal{T}$ is a rooted tree). That is, if we denote the length of an edge e of $\mathcal{T}$ by λ_e we have:

$$PD(Y) = \sum_e \lambda_e,$$

where the summation is over all edges e in $\mathcal{T}$ that lie on the minimal subtree of $\mathcal{T}$ connecting the taxa in Y (and if $\mathcal{T}$ is rooted, also connecting the root). There has been some debate about whether the root should be included, however the original definition in [14] and prevailing usage include the root (see [10], [15] and [9] for further discussion). Figure 6.1 illustrates the various PD measures we have discussed here.

Depending on the data from which a tree is derived, the branch lengths may have different interpretations. Branch lengths may correspond to an evolutionary time-scale (i.e. the number of millions of years between speciation events), or to genetic distance, or to the extent of morphological differences, or perhaps some combination of these (or other) measures of evolutionary distance. Throughout this chapter, no particular interpretation is assumed, so as to allow the greatest degree of generality for applications; in particular, unless we state so explicitly, we do not assume that the tree is ultrametric (an *ultrametric tree* is one for which the distance from the root to any leaf is the same, as would occur for (a) genetic distance under a 'molecular clock', or (b) an evolutionary time-scale).

The PD measure has two basic combinatorial properties which we now describe.

6.2.1 *The strong exchange property*

For any function f defined from the collection of subsets of X of size at least r into the real numbers, we say that f satisfies the *strong exchange property* if for any two subsets Y and Z with $r \leq |Y| < |Z|$ there exists some taxon $z \in Z - Y$ such that:

$$f(Z - \{z\}) - f(Z) + f(Y \cup \{z\}) - f(Y) \geq 0. \tag{6.1}$$

This condition is a sufficient condition for the greedy algorithm to construct subsets of any given size ($\geq r$) that maximize f starting from any given set of size r that maximizes f. This follows by standard arguments from 'greedoid' theory (see [23]). To construct such a subset the greedy algorithm iteratively adds the element (taxon) that gives a maximal increase in f until the subset contains r elements.

The strong exchange property was established for $f = PD$ (and $r = 2$ in the case of unrooted phylogenetic trees) in [46]; its interpretation in this setting is that for any two of the subsets, the larger one contains some taxon (z) that would contribute at least as much to the PD value of the smaller subset than it adds to that of the larger one.

Consider both trees in Fig. 6.1 and the situation where the two subsets Y and Z are $\{a, c\}$ and $\{b, d, e\}$ respectively; clearly $|Y| < |Z|$. Deleting taxon b from subset Z and adding it to Y results in a loss of the combined PD of Y and Z in both trees, hence b does not satisfy the strong exchange property. However the combined PD of Y and Z is increased if taxon d is removed from Z and added to Y, thus satisfying the strong exchange property.

Note that the strong exchange property for PD fails for $r = 1$ and $r = 0$ for unrooted trees, but holds for rooted trees. Moreover, as demonstrated in [34], for any given set of taxa W of size at least 2 (or 1 in case of rooted trees) the strong exchange property also ensures that amongst the collection of all subsets of size k containing W, the one(s) of maximal PD value can be constructed from W by the greedy algorithm (even though W itself may not have optimal PD score for its cardinality).

6.2.2 *Generalized Pauplin formula*

The second combinatorial property of PD is that it can be written canonically as a linear combination of pairwise distances within the tree. That is, if $d(x, y)$ denotes the distance between x and y in $\mathcal{T}$, the PD of a set W can be written as

$$PD(W) = \sum_{x,y \subseteq W} \mu_{\mathcal{T},W}(x, y) d(x, y) \tag{6.2}$$

where $\mu_{\mathcal{T},W}$ is a function that depends on $\mathcal{T}$ and W but not the branch lengths. Actually there are many possible choices of $\mu_{\mathcal{T},W}$ but there is one that is particularly natural and which is defined as follows. Let $\mathcal{T}_W$ denote the subtree of $\mathcal{T}$

connecting W and let $p(\mathcal{T}_W, x, y)$ be the set of non-leaf vertices of $\mathcal{T}_W$ that lie on the path connecting x and y. Then set

$$\mu_{\mathcal{T},W}(x,y) = \prod_{v \in p(\mathcal{T}_W, x, y)} (d(v) - 1)^{-1}$$

where $d(v)$ is the degree of vertex v in $\mathcal{T}_W$. The validity of equation (6.2) for this choice of $\mu_{\mathcal{T},W}$ was described (for $W = X$) for binary phylogenetic X–trees by Pauplin [35], and generalized to arbitrary phylogenetic X–trees in Semple and Steel [43]. The Pauplin formula also provides an interesting starting point for forming species specific indices of biodiversity such as the Equal-Splits index (Section 6.3.1).

6.2.3 *Exclusive molecular phylodiversity*

We end this section by noting that Lewis and Lewis [26] recently investigated a related measure they called the 'exclusive molecular phylodiversity' of a set Y, defined by:

$$E(Y) = PD(X) - PD(X - Y).$$

This measure has also been used by [41] to assess the evolutionary history of endemic species in biodiversity hotspots. The benefit of exclusive molecular phylodiversity in that context is that it avoids the need for any information about non-endemic species, effectively assuming that these are well represented elsewhere. It is easy to show that this measure does not satisfy the strong exchange property (equation (6.1)) and that greedy algorithms cannot be guaranteed to produce an optimal subset, Y.

6.3 Biodiversity conservation

Ross Crozier summarizes the rationales for conserving biodiversity into three categories: 'moral (other species have a right to exist), esthetic (species are like works of art, and it would be foolish to destroy them), and utilitarian (humans derive material benefit from the existence of other species)' [8]; these motivations are further explored in [31]. Given unlimited resources for conservation all three motivations dictate the same action—conserving all taxa. In a realistic setting where there are limited resources for conservation the taxa must be prioritized in some manner. In this case the three categories of motivation may dictate different prioritizations.

If conservation is motivated by moral considerations, as many taxa as possible should be conserved. A conservation scheme should therefore allocate its resources so that the net survival increase of all taxa is as high as possible.

If the motivation for conservation is utilitarian, the distinctiveness of the remaining taxa is of great importance. For example, protecting the sole remaining taxon from a clade has greater utilitarian benefits than protecting a taxon from a well represented clade as the former has greater unique genetic potential for further evolution and bio-prospecting [8].

Lastly, if conservation is motivated by aesthetic reasons the role that distinctiveness should play is dependent on the uncertain definition of aesthetic value. However, given the choice of saving either a taxon from a well represented clade or a taxon that is the 'last of its kind' it seems difficult to find a general justification for not choosing the latter.

Most biodiversity conservation approaches aim to conserve as many taxa as possible [18], but the reasons used to motivate conservation are often utilitarian in nature (e.g. Chapter 1, [37]) and should therefore take taxon distinctiveness into account.

In this subsection we discuss several methods for prioritizing taxa to conserve that allow distinctiveness to be taken into consideration. Throughout this chapter we apply various methods and indices to the tree depicted in Fig. 6.2. This tree shows the phylogenetic relationship of Crested penguins (*Eudyptes*) as produced by Sara Bertelli and Norberto Giannini [3],[19]. Note that no edge lengths were given in the original tree, and so here we have assumed that the tree is ultrametric, as shown in Fig. 6.2. We are using this example for illustrative purposes only, but it is of interest to note that according to the 2004 IUCN Red

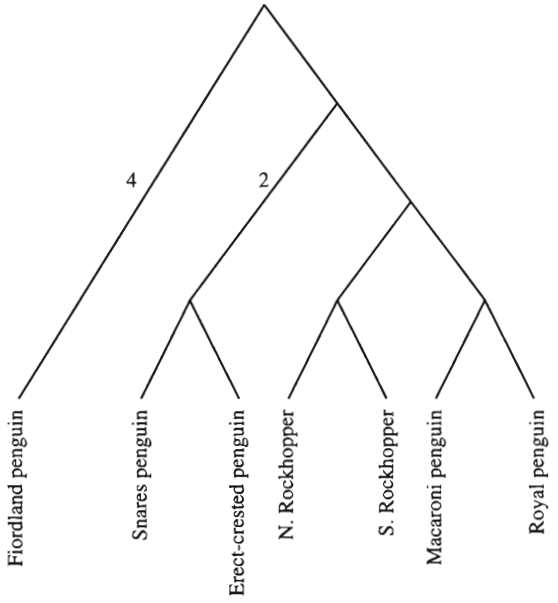


FIG. 6.2. The phylogenetic tree for Crested penguins. This tree was derived from the tree in [3] and [19] which had no branch lengths. For illustrative purposes each level in the original tree was assumed to be separated by the same distance such that all edges in this tree are of length 1 except for the two marked edges.

List [22] all of the species are vulnerable except the Erect-crested penguin, which is endangered.

6.3.1 *Simple indices*

Many indices for measuring or ranking the distinctives of a single taxon have been proposed. These indices have the advantage of being easy to compute and, as each taxon is assigned some value, they can be readily combined with other information for decision making (such as the conservation cost or economic importance of that taxon). The disadvantage of these indices is that they do not take into account the complexities of conserving multiple taxa. For example, if one taxon is conserved the relative importance of conserving closely related taxa may decrease as we wish to conserve as distinctive a *set* of taxa as possible.

Here we consider some simple indices with a particular focus on those that are based on the concept of phylogenetic diversity. For the interested reader some notable work not considered here (some of which is not based on PD) is contained in [2], [6], [7], [20], and [51].

One of the conceptually simplest indices is the 'Pendant Edge' (PE) measure introduced by Stephen Altschul and David Lipman [1] where each taxon is assigned a value equal to the length of its pendant edge. Strictly speaking the PE measure is not based on PD but has been included here for illustrative purposes.

The PE value of each species in Fig. 6.2, is easily determined: in this case the Fiordland penguin has the highest PE with the other taxa having an equal second highest value. PE suggests that the Fiordland penguin is the most important to conserve but does not differentiate between the other taxa. It seems logical, though, that if some of the other taxa were to be conserved, we should not choose the most closely related of these.

A conceptually appealing family of indices divides the total phylogenetic diversity of a tree amongst the taxa corresponding to the leaves of that tree. An example of these indices is the Equal-Splits (ES) index [38] which is closely related to the previous discussed Pauplin formula. This index splits the PD value of an edge equally between it's daughter trees. Denoting the edge length between a node, j, and its direct ancestor by λ_j, the equal splits index for a taxon, i, can be calculated by summation over all the nodes between i and the root (including i):

$$ES(i) = \sum_j \frac{\lambda_j}{2^{d'(i,j)}}$$

where $d'(i,j)$ is the number of edges between the taxon (node i) and node j. Applying the ES index to the tree in Fig. 6.2 again suggests that the Fiordland penguins are the most important species to conserve with an index value of 4. The Snares and Erect-crested penguins have an index equal value of $\frac{9}{4}$ whilst the remaining species have a value of $\frac{15}{8}$; if, for example, three species could be conserved, this suggests that the Fiordland, Snares, and Erect-crested penguins

should be chosen. Intuitively, however, it seems more beneficial to conserve one of the other species instead of the Snares or Erect-crested penguins and thus protect more of the internal edges. The problem with ES (and other simple indices) is that the decision to conserve one taxon does not affect the importance assigned to conserving the remaining taxa.

One can also consider the expected contribution to PD that a taxon will make at some time in the future if the survival of all other taxa is uncertain. To make this idea precise, for each subset S of X, and each taxon $i \in X - S$ let

$$\Delta_{PD}(S, i) = PD(S \cup \{i\}) - PD(S);$$

$\Delta_{PD}(S, i)$ is the increase in PD that taxon i provides when added to S. Now, suppose that each taxon $j \in X - \{i\}$ has a probability a_j that it is not extinct at some time t in the future. If we assume that extinction events are independent between taxa, and let E be the (random) set of taxa that are extant at time t then we can ask how much we expect taxon i to contribute to the PD at time t—this value, ψ_i is simply the expected value of $\Delta_{PD}(E, i)$, given formally by

$$\psi_i = \sum_{S \subseteq X - \{i\}} \mathbb{P}[E = S] \Delta_{PD}(S, i).$$

Note that $\mathbb{P}[E = S]$ is the probability that the set of extant taxa at time t will be S; this depends on the survival probabilities (a_j's).

Although this last equation involves a summation over an exponential number of terms, it has an equivalent description that allows for its rapid (polynomial-time) calculation (Steel, M., A. Mimoto and A. O. Mooers, submitted). A related but different index to ψ_i is the Shapley value which has been considered in detail elsewhere [20].

6.3.2 *Noah's Ark Problem*

Martin Weitzman introduced the 'Noah's Ark Problem' (NAP) [52], a comprehensive framework for allocating limited funding for biodiversity conservation that overcomes some of the problems associated with the simple indices discussed previously. In the NAP framework each taxon, j, has some probability, a_j, of remaining extant. If some conservation intervention of cost c_j is applied to this taxon, then this survival probability can be increased from a_j to b_j. The aim is to identify the subset of taxa to conserve: S, that maximizes the future expected phylogenetic diversity $\mathbb{E}(PD|S)$ subject to the budgetary constraint, B. The notation '$|S$' indicates that the expected value is conditional on the set S of taxa being conserved. The formulation of the NAP as used in this chapter is:

Given an edge-weighted phylogenetic tree, and values (a_j, b_j, c_j) *for each taxon* j, *maximize* $\mathbb{E}(PD|S)$ *over all subsets* S *of taxa, subject to the constraint:* $\sum_{j \in S} c_j \leq B$.

The original formulation used by Weitzman [52] allowed the inclusion of an intrinsic value for the taxa (for example the tourism value of a species of whale), however this value can readily be included here by increasing the length of each taxon's pendant edge appropriately. There is of course an inherent difficulty in combining phylogenetic diversity and other socio-economic values; this is also the case in the original formulation of the NAP.

$\mathbb{E}(PD|S)$ is calculated by summing all the edge lengths, λ_e, in the tree, weighting each edge by the probability that it will be spanned by the surviving taxa:

$$\mathbb{E}(PD|S) = \sum_e \lambda_e p(e|S).$$

For rooted trees the probability that an edge is spanned, $p(e|S)$, is simply the probability that at least one of the taxa in the tree subtended by edge e will remain extant.

Variations of the NAP have been used in a variety of applications including biodiversity conservation (e.g. [10], [29], and [45]) and prioritizing taxa for genomic sequencing [34]. Additional intrinsic values for the taxa can be incorporated in this version of the NAP by adding the intrinsic value of each taxon to its pendant edge.

A problem with the NAP is that no efficient algorithm has been found for producing solutions to it. To find an optimal solution it may be necessary to consider many of the possible subsets of taxa. The number of subsets increases at rate $2^{|X|}$, therefore considering a large proportion of these is infeasible for more than a few dozen taxa. For example, if one has a tree with (say) 1,000 taxa, and one wishes to find a subset of (say) 100 taxa that maximizes $\mathbb{E}(PD|S)$ then it is impossible for any computer to search all subsets of size 100 from the 1,000. Having efficient algorithms for solving the NAP is therefore essential for applying the NAP to large trees.

Several variations of the NAP where additional constraints are imposed have been shown to be solvable using simple 'greedy' algorithms [21], [46]. These algorithms allow the optimal solutions for a particular problem to be found quickly. Here we provide a further extension to the scenario considered in [46].

First consider the class of NAPs where taxa become extinct unless they are conserved, all taxa cost the same to conserve and conserved taxa survive with certainty; this corresponds to $a_j = 0, b_j = 1$, and $c_j = c$ (where $c > 0$ is some constant) for each taxon j. We will call this type of NAP *Scenario 1.*

In this scenario, the expected remaining phylogenetic diversity ($\mathbb{E}(PD|S)$) is simply the phylogenetic diversity of the conserved taxa ($PD(S)$), since all other taxa become extinct with certainty. Solving the NAP is therefore equivalent to finding the subset S of X of size at most $\frac{B}{c}$ with maximal PD. This problem was shown to be solvable using a simple greedy algorithm in [46], from which we have the following result:

Theorem 6.1 *For a NAP under Scenario 1, the following greedy algorithm produces the optimal solution(s). For rooted trees the algorithm begins with an*

empty set S, and for unrooted trees it begins with a set S containing the two taxa that are furthest apart. The algorithm sequentially adds the taxon that provides the greatest increase in $\mathbb{E}(PD|S)$ until S contains as many taxa as the budget permits to be conserved. Where more than one taxon provides an equal increase in $\mathbb{E}(PD|S)$ one is chosen at random. Upon completion S contains an optimal solution, other optimal solutions (if they exist) are obtained by making different choices where a taxon was chosen at random.

We will now extend Scenario 1 to allow non-zero survival probabilities in the absence of conservation ($a_j \neq 0$), as follows. We will refer to this extension as *Scenario 2* which has the remaining constraints that $b_j = 1$, c_j is constant and the tree is rooted. The following result was independently derived here and in [33].

Theorem 6.2 *For a NAP under Scenario 2, the greedy algorithm described in Theorem 6.1 produces the optimal solution(s) when applied to a rooted tree with suitably adjusted edge lengths, λ'_e. Denoting the set of children of edge e (the leaves/taxa separated from the root by e) by C_e the adjusted edge lengths are:*

$$\lambda'_e = \lambda_e \prod_{j \in C_e} (1 - a_j). \tag{6.3}$$

Proof Instead of maximizing $\mathbb{E}(PD|S)$ we can seek to maximize $\mathbb{E}(PD|S) - \mathbb{E}(PD|\emptyset)$, the increase in the expected PD that conservation of the taxa in S will provide. For a Scenario 2 problem the increase in the probability that a particular edge is spanned when the set, S, of taxa is conserved is:

$$\begin{aligned} p(e|S) - p(e|\emptyset) &= \begin{cases} 1 - (1 - \prod_{j \in C_e}(1 - a_j)), & \text{if } |C_e \cap S| > 0; \\ 0, & \text{if } |C_e \cap S| = 0; \end{cases} \\ &= \prod_{j \in C_e} (1 - a_j) \times \begin{cases} 1, & |C_e \cap S| > 0; \\ 0, & |C_e \cap S| = 0. \end{cases} \end{aligned}$$

The expected increase in the PD is simply the sum over all edges with each edge weighted by the increased probability:

$$\begin{aligned} \mathbb{E}(PD|S) - \mathbb{E}(PD|\emptyset) &= \sum_e \lambda_e (p(e|S) - p(e|\emptyset)) \\ &= \sum_e \lambda_e \prod_{j \in C_e} (1 - a_j) \times \begin{cases} 1, & \text{if } |C_e \cap S| > 0; \\ 0, & \text{if } |C_e \cap S| = 0; \end{cases} \\ &= \sum_e \lambda'_e \times \begin{cases} 1, & \text{if } |C_e \cap S| > 0; \\ 0, & \text{if } |C_e \cap S| = 0. \end{cases} \end{aligned}$$

This final expression for $\mathbb{E}(PD|S) - \mathbb{E}(PD|\emptyset)$ is equal to the objective, $\mathbb{E}(PD|S)$, for a Scenario 1 problem with branch lengths λ'_e as required. □

6.3.3 *Conservation time scale*

The survival probabilities (a_j) contain an implicit time scale as they represent the probability that a taxon will survive to some future time, t; in the absence of conservation the expected number of taxa surviving to t is $\sum_j a_j$. If the time t is in the distant future (a long time scale) the survival probability of unprotected taxa will be close to zero due to background extinction, for shorter time scales (t closer to the present) the survival probabilities will be closer to one. This choice of time scale affects solutions to the NAP as management strategies corresponding to longer time scales will place greater emphasis on internal edges. Note that Scenario 1 corresponds to long term management where only those taxa that were conserved remain, whereas in Scenario 2 the time scale can be freely chosen by selecting values for a_j that are of appropriate magnitude.

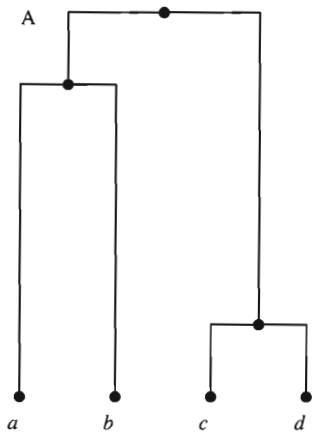
To illustrate the importance of selecting an appropriate time scale consider the tree in Fig. 6.3, where each taxon is equally likely to remain extant at any future time. Panel A corresponds to the situation where all taxa that are not conserved become extinct (a long time scale). If two taxa can be conserved, the optimal choice consists of one taxon from each branch of the tree. This optimal choice is found either by application of the greedy algorithm (Theorem 6.1) or by an exhaustive search.

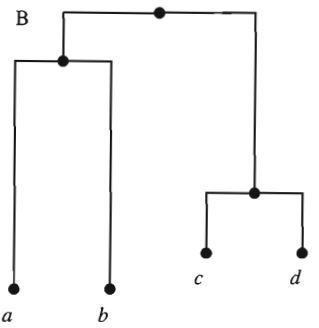
Consider increasing the survival probability of unconserved taxa (a_j) so that all taxa have a $\frac{1}{4}$ chance of surviving; this represents a move to a shorter management time scale. To find the optimal solutions for this problem the transformation outlined in Theorem 6.2 is applied to the original tree (Panel A in Fig. 6.3) yielding the tree in Panel B. As expected from equation (6.3) the interior edges have had a greater reduction in length than the pendant edges; application of the greedy algorithm can now be used to obtain the optimal solutions. The pendant edge lengths of taxa a and b are now equal to the distance between the root and taxa c or d. Consequently conserving both taxa a and b is now also an equally good solution.

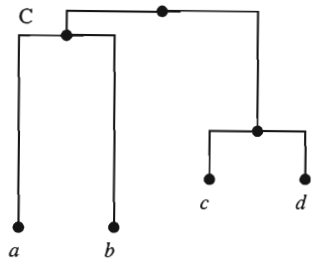
If the survival probabilities (a_j) are further increased (to, say, $\frac{3}{8}$), the interior edges of the transformed tree decrease in length to such an extent that the optimal set of taxa to conserve becomes $\{a, b\}$ (see Panel C).

We have illustrated that the optimal set of taxa to conserve is dependent on the management time scale. As the management time scale shifts from long term to short term, less emphasis is placed on interior edges as these are more likely to remain extant anyway.

A discussion of the merits of conservation time scales is beyond the scope of this work (see [4] and [25] for more details). However the optimal time scale will be highly dependent on the application. Of particular importance will be the time scale on which conservation focus can be shifted from one taxon to another. If this can occur rapidly, planning for the short term would be optimal and the conservation strategy should be reevaluated as taxa become extinct. For







Conserved Taxa (S)	Optimal? A	B	C
$\{a, b\}$	✗	✓	✓
$\{c, d\}$	✗	✗	✗
$\{a, c\}$,$\{a, d\}$, $\{b, c\}$,$\{b; d\}$	✓	✓	✗

FIG. 6.3. Panel A depicts a tree where unconserved species become extinct with certainty ($a_j = 0$). Panels B and C depict the transformed tree as this survival probability is increased to 0.25 and 0.375 respectively. Optimal subsets of size 2 can be found by applying the greedy algorithm to these trees. The optimality of each subset for each panel is indicated in the table.

many taxa, conservation programmes are long term investments. In these cases, a longer time scale should be investigated when the taxa to be conserved are initially selected.

6.3.4 *Further algorithmic results*

For problems where a greedy algorithm is known to produce optimal solutions, a naïve implementation of the algorithm may be unnecessarily slow. An efficient implementation of the greedy algorithm for Scenario 1 is provided in [28], which in their simulations took 1/100th of the time of a naïve implementation. In [28] an alternative pruning algorithm is also provided, this algorithm begins with all the taxa and removes the least important taxon sequentially until a subset of the desired size is obtained. As expected if a large proportion of the taxa are to be included in the subset, the pruning algorithm is more efficient.

Two further variations of the NAP for which greedy algorithms produce optimal solutions were considered by the authors in [21]. The first variation permits the survival probability for conserved and unconserved taxa (a_j and b_j) to be varied, but these must be related by a particular relationship. The second variation permits variable conservation costs (c_j) but requires that taxa only survive if they are conserved ($a_j = 0$, $b_j = 1$). Additionally, for the greedy algorithm to produce optimal solutions, the tree must be ultrametric (satisfy a molecular clock).

A dynamic programming algorithm has also been produced for a less restrictive variation of the NAP with the sole restriction that conserved taxa survive with certainty ($b_j = 1$) [33].

6.3.5 *Extensions to the NAP*

The Noah's Ark Problem provides a satisfying framework for biodiversity resource allocation problems. It is, however, still a simplification of reality and some extensions to it have been suggested.

The NAP as presented here does not consider the possibility of partially conserving taxa and therefore being able to spread resources more thinly across a greater number of taxa. Weitzman [52] assumed that the survival probability of a taxon increases linearly with the conservation funding allocated to that taxon. Under this assumption optimal solutions to the NAP are extreme and allocate the maximum possible amount to a few taxa instead of partially conserving a greater number. An extension of the NAP to more realistic relationships between survival probability and expenditure was considered in [44], with an application to conservation of breed diversity in African cattle. A greedy algorithm was presented in that paper that the authors suggested would provide optimal solutions to all problems of this type. However, it was shown in [21] that this cannot be the case. This was extended further in [39] to allow for discontinuous relationships produced by multiple possible conservation schemes, necessitating a two step optimization procedure (which they state is not guaranteed to produce the global optimum).

Another implicit assumption in the NAP is that the survival probabilities are independent. That is, conserving one taxon does not raise or lower the survival probabilities of any others, and this may be unrealistic. For example, conserving the prey of one taxon may raise the survival probability of that taxon as well. This effect was considered in [50] where it was shown that failure to consider interdependent survival probabilities may result in an incorrect suggestion as to which species should be protected. The authors in this study stress the importance of their findings as 'more significant losses of biodiversity are exactly those in which ecological impacts are severe, that is, where the loss of one species affects the survival of others'.

In summary, whilst the NAP provides a good starting point, there are other important factors that influence which taxa should be conserved. Inclusion of some of these may prove more difficult than others and adding these factors will further complicate the problem of finding optimal solutions. For example,

consider the following problem which is relevant to biodiversity conservation. We have a collection C of locations, where each location $l \in C$ contains some subset $S(l)$ of taxa from a set X of taxa; also we have a phylogenetic X–tree $\mathcal{T}$ with branch lengths. We wish to select k locations so as to maximize the PD of the set of taxa that occur in at least one selected location. If no taxon occurs in more than one location this problem is easily solved, by transforming it to the standard PD optimization problem and applying the greedy algorithm. In general, however, the problem is NP-hard. The proof consists of showing that one can transform the NP-complete problem 'Minimum cover' [16] to this problem, by selecting branch lengths for $\mathcal{T}$ that are 1 on all the pendant edges, and 0 on all the interior edges. For various approaches to solving this and related problems see [40], [5] and [53].

6.4 Loss of phylogenetic diversity under extinction models

We turn now to the statistical properties of PD as taxa go extinct, beginning with a recent result from [47]. Nee and May [30] investigated the loss of PD as taxa are randomly deleted from random trees under a simple model: each taxon is equally likely to be the next to become extinct (the 'field of bullets' model). The trees were ultrametric trees as generated by a random-birth model. They found a characteristic concave shape in the relationship between the expected remaining PD and the proportion of taxa deleted. This relationship is illustrated for the Crested penguins tree (Fig. 6.2) by the upper curve in Fig. 6.4.

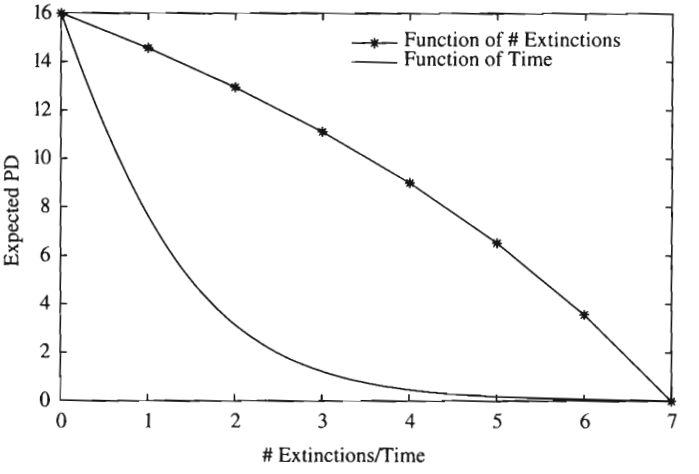


FIG. 6.4. The expected remaining PD after extinctions have occurred among the Crested penguins depicted in Fig. 6.2. This loss in PD is viewed as a function of both the number of extinctions that have occurred and the time that has elapsed since extinctions have been allowed to occur.

This relationship was further investigated recently in [45], which studied random deletion of taxa from certain biological trees. Once again the relationship between taxa deleted and remaining PD was concave. Recall that a sequence $x = (x_1, x_2, \ldots, x_n)$ of real numbers is *concave* if, when we let $\Delta x_r = x_r - x_{r-1}$ the following inequality holds for all r:

$$\Delta x_r - \Delta x_{r+1} \geq 0$$

and the sequence is *strictly concave* if the inequality is strict for all r. Geometrically this means that the slope of the line joining adjacent points in the graph of x_r versus r is decreasing. Note that x_r is concave precisely if the complementary (reverse) sequence $y_r = x_{n-r}$ is concave. The significance of (strict) concavity for PD is that it says (informally) that most PD loss comes near the end of an extinction process.

In this section we first describe a generic concave relationship observed between the average PD and the number of taxa deleted. This makes intuitive sense, because each interior branch survives until the point where there is no taxon below it and this is likely to occur towards the end of a random extinction process.

Consider a rooted phylogenetic tree having a leaf set X of size n. Let W be a random subset of taxa of size r sampled uniformly from X (for example, by selecting uniformly at random a set S of $n - r \geq 0$ elements of X and deleting them, in which case $W = X - S$). For $r \in \{1, \ldots, n\}$ let $\mu_r = \mathbb{E}[PD|r]$, the expected value of $PD(W)$ over all such choices of W. Equivalently, we can write $\mu_r = \binom{n}{r}^{-1} \sum_{W \subseteq X : |W|=r} PD(W)$, where $\binom{n}{r}$ is the binomial coefficient $(= \frac{n!}{r!(n-r)!})$, which is the number of ways of selecting r elements from a set of size n. For brevity we adopt the usual convention that $\binom{n}{r} = 0$ if r is greater than n or less than 0.

Clearly $\mu_n = PD(X)$. For $r \in \{1, \ldots, n\}$, let $\Delta\mu_r = \mu_r - \mu_{r-1}$. Note that, since $\mu_0 = 0$, we have $\Delta\mu_1 = \mu_1$. For an edge e of $\mathcal{T}$, and $r \in \{1, \ldots, n-1\}$ let

$$\psi(e, r) := \frac{n_e(n_e - 1)}{r(r+1)} \cdot \frac{\binom{n-n_e}{r-1}}{\binom{n}{r+1}}$$

where n_e denotes the number of leaves of $\mathcal{T}$ that lie 'below' e (i.e. separated from the root by e).

The proof of the following result is given in [47]. It shows that for any fully resolved tree, PD decays in a strictly concave fashion as taxa are randomly deleted, and the only trees for which the decay of PD is linear are fully unresolved 'star' trees. In the following theorem a *cherry* is a pair of leaves that are adjacent to the same vertex.

Theorem 6.3 *Consider a phylogenetic tree $\mathcal{T}$ with an assignment λ of positive branch lengths. Then, for each $r \in \{1, \ldots, n-1\}$,*

$$\Delta\mu_r - \Delta\mu_{r+1} = \sum_e \lambda_e \psi(e, r)$$

where the summation is over all edges of $\mathcal{T}$. In particular, μ is concave over this domain, and μ is strictly concave if and only if $\mathcal{T}$ has a cherry, while μ is linear if and only if $\mathcal{T}$ has no interior edges (i.e. is an unresolved 'star' tree).

Consider the tree for Crested penguins to which we have previously referred (Fig. 6.2). Figure 6.4 shows the expected PD as a function of the number of extinctions. As expected from the above theorem, the relationship depicted in this figure is strictly concave.

6.4.1 *Relationship between PD and time under an extinction process*

We have investigated the expected PD as a function of the number of extinctions that have occurred. So far each taxon has been considered as equally likely to be the next to become extinct. However, no consideration has been given to the timing of these extinctions. Here we consider the situation where each taxon has the same probability of becoming extinct at any point in time (the time to extinction for an individual taxon has an exponential distribution) and consider the expected PD as a function of the time instead of the number of extinctions that have occurred. We will show that the decline in expected PD does not in general have a concave shape and in fact after a specific time (dependent on the tree shape) the decline will become convex. Note that this is not a contradiction with the previous result; it is simply due to the fact that the number of extinctions decreases over time as there are fewer species left that could become extinct.

The probability that an edge, e, will be spanned by the taxa remaining at some time t, depends only on the number of children ($|C_e| = n_e$) of that edge. Denoting this probability by $p_e(t)$ we have:

$$p_e(t) = 1 - \left(1 - e^{-rt}\right)^{n_e}$$

where r is the rate of extinction. The expected PD at time t, $\mathbb{E}_t(PD)$ is easily found using these probabilities:

$$\mathbb{E}_t(PD) = \sum_e \lambda_e p_e(t).$$

Observe that $\mathbb{E}_t(PD)$ depends only on the sums of the edges with the same number of leaves attached, not on the individual edges themselves:

$$\mathbb{E}_t(PD) = \sum_{j=1}^{m} \alpha_j \left[1 - \left(1 - e^{-rt}\right)^j\right],$$

where $\alpha_j = \sum_{e, n_e = j} \lambda_e$, and m is the highest number of leaves below any edge—this corresponds to the edge(s) at the root with the most leaves descendant from them. To investigate the shape of $\mathbb{E}_t(PD)$ the second derivative is easily obtained:

$$\frac{d^2\mathbb{E}_t(PD)}{dt^2} = r^2 e^{-rt}\left(\alpha_1 + \sum_{j=2}^{m} \alpha_j j \left(1 - je^{-rt}\right)\left(1 - e^{-rt}\right)^{j-2}\right). \qquad (6.4)$$

For convexity, the second derivative must be positive. The term corresponding to α_1 is clearly positive, but the sign corresponding to the other α-values depends on t. The term corresponding to a particular α_j is positive if $1 - je^{-rt} > 0$ which holds when

$$t > \frac{\ln(j)}{r}.$$

A sum of convex functions is convex, therefore once the above condition is satisfied for all j, $\mathbb{E}_t(PD)$ will be convex. The term that becomes convex the latest is the term with the highest value of j (namely m). Convexity is therefore guaranteed after $\hat{t} = \ln(m)/r$. In the limit as $\sum_{j<m} \alpha_j/\alpha_m \to 0$, $PD(t)$ will become convex exactly at $\hat{t}$, however $PD(t)$ will generally become convex earlier due to the other terms.

The terms corresponding to edges with high values of j are the last to become positive; as more weight is assigned to these the time to convexity lengthens. Variation in diversification rates through time and/or among clades can therefore affect the time to convexity.

The amount of PD loss that has occurred by the time that convexity is guaranteed ($\hat{t} = \ln(m)/r$) is difficult to characterize, but the number of taxa remaining at this time can be readily found. The probability of an individual taxon persisting to time t is e^{-rt}, so at $t = \hat{t}$ each taxon is extant with probability $1/m$. The total number of taxa is between $m + 1$ and $2m$ (depending on the imbalance of the tree at the root) and the expected number of extant taxa at $t = \hat{t}$ is therefore between 1 and 2. Accordingly, the convexity result may appear to be of limited biological interest, however, given a real tree, the expected number of taxa remaining by the time convexity is reached will usually be much higher.

Another interesting behaviour that can readily be examined and may be of more practical interest is the initial shape of the PD decline (that is at and just after $t = 0$). Substituting $t = 0$ in equation (6.4) we obtain:

$$\begin{aligned}\frac{d^2\mathbb{E}_t(PD)}{dt^2}|_{t=0} &= r^2\left(\alpha_1 + \sum_{j=2}^{m}\left(\alpha_j j\,(1-j)\,0^{j-2}\right)\right)\\ &= r^2\,(\alpha_1 - 2\alpha_2)\,. \qquad (6.5)\end{aligned}$$

Initial convexity requires $\alpha_1 > 2\alpha_2$ and concavity requires $\alpha_1 < 2\alpha_2$. The edges that contribute to α_1 are the pendant edges and those contributing to α_2 are edges above cherries. Any tree can have at most half as many 'above cherry' edges as pendant edges, so if pendant edges have similar lengths as the 'above cherry' edges then that tree will therefore exhibit initial convexity (as for the Crested penguins tree Fig. 6.2 and 6.4). It should be noted that even if the PD loss curve for a tree is convex at $t = 0$ and after $t = \hat{t}$ there is no guarantee that it will be convex between these two times due to the complexity of equation (6.4).

6.5 Tree reconstruction using PD

The simplest form of PD (on unrooted trees) considers subsets of taxa of size 2, in which case the PD value is just the path distance in the tree connecting the two taxa. Such pairwise distances suffice to reconstruct any tree (and indeed also the branch lengths). This is a classic result dating back to the mid-1960s [54], and it forms the basis of many fast and popular tree-building methods, such as Neighbor-Joining and BioNJ. However, despite their usefulness, pairwise distances have some drawbacks, and in this section we explore some of the ways in which PD–values on subsets of m–taxa (for $m > 2$) may provide a promising approach in future.

One (statistical) concern with using pairwise distance data is that converting sequence data to pairwise distances is a highly reductive transformation. That is, each distance matrix typically can be obtained from a huge number of different sets of aligned sequences, even under the usual Hamming distance measure (and even if we just count the frequencies of site patterns, not the order they occur in, [48]). Whether this extensive 'loss of information' is important for phylogeny reconstruction is a tantalizing question, though it is tempting to conjecture that it is. Phylogenetic diversity is one way of generalizing the idea of a distance in a tree—from pairs of leaves, to m-tuples of leaves—and this measure suggests a natural way of refining distance-based approaches, so that less information is lost in using sequences to build trees.

To illustrate this idea, consider a model-based approach to phylogeny reconstruction. Given a model of sequence evolution, one can generally compute the maximum-likelihood estimate of an 'evolutionary distance' $d(x, y)$ between any two sequences x, y. This 'evolutionary distance' is some quantity that is assumed to be additive on the underlying evolutionary tree. For example, for a stationary reversible Markov process of site substitution, the 'evolutionary distance' between x and y is usually understood as the expected number of substitutions occurring on the path separating x and y. Thus $d(x, y)$ can be viewed as an estimate of $PD(\{x, y\})$ for a suitable edge weighting of $\mathcal{T}$.

Notice that the PD values on subsets of X of size 3 are determined by the pairwise PD values, according to the following 3–point condition:

$$2PD(\{x, y, z\}) = PD(\{x, y\}) + PD(\{y, z\}) + PD(\{z, x\}). \tag{6.6}$$

Thus one could estimate $PD(\{x, y, z\})$ by using the pairwise distance estimates d, but again this results in a loss of information in reducing triplewise data to three pairwise marginals. Thus it may be more appropriate to estimate PD on m-element subsets by direct analysis of sequence data. For example, the PD score for three sequences might be estimated as the sum of the three branch lengths that maximize the likelihood score of the three sequences under a Markov process of site substitution (and perhaps also insertion and deletion). For certain models, the PD value when $m = 3$ can also be calculated explicitly (i.e. without optimizing branch lengths to maximize likelihood) by the 'tangle' triplewise distance described in [49].

When $m = 3$, estimation of PD values does not require estimating the tree structure connecting the m taxa. However, for any value $m > 3$, consideration of different trees connecting the m taxa is necessary.

Suppose that one was able to exactly calculate the true PD values for all m-elements subsets of X. A natural question is whether this information uniquely determines the underlying phylogenetic X-tree $\mathcal{T}$. It is clear that in general the answer is 'no'—for if we take $m = |X|$ then we have just one PD value, and this can be realized on any phylogenetic X-tree by taking appropriate branch lengths. However, Pachter and Speyer [32] recently showed that if m does not exceed $(n+1)/2$ then the tree $\mathcal{T}$ is uniquely determined by the PD scores of the m-element subsets of X. More precisely, their result states:

Theorem 6.4 *Let $\mathcal{T}$ be a phylogenetic X-tree (with $n = |X|$) and $m \geq 2$ an integer. If $n \geq 2m - 1$ then $\mathcal{T}$ is determined by the map that associates each m-element subset of X with its induced PD score.*

Moreover, even when m exceeds $(n+1)/2$ some partial information concerning $\mathcal{T}$ can be recovered from this map [24]. This paper also describes a modification of Neighbor-Joining to reconstruct trees from their induced PD values. The central idea here is to identify a cherry of the tree. The following result (the 'cherry-picking theorem' of [24]) generalizes the way that Neighbor-Joining identifies cherries in the special case $m = 2$.

Theorem 6.5 *Suppose that $\mathcal{T}$ is a phylogenetic X-tree with n leaves, and m is any integer between 2 and $n - 2$. Then any distinct pair $i, j \in X$ that minimizes the expression*

$$\binom{n-2}{m-1} \sum_{\substack{Y \subset X: \\ i,j \in Y, |Y|=m}} PD(Y) - \sum_{\substack{Y \subset X: \\ i \in Y, |Y|=m}} PD(Y) - \sum_{\substack{Y \subset X: \\ j \in Y, |Y|=m}} PD(Y)$$

is a cherry of $\mathcal{T}$.

Phylogenetic diversity also forms the basis of other approaches to tree reconstruction—most notably the 'balanced minimum evolution' (BME) method of Pauplin [35]. This method takes a (pairwise) distance estimate d on X as input and scores each resolved phylogenetic X-tree $\mathcal{T}$ by what d would estimate for $PD(X)$ using equation (6.2). Thus, if d is additive on $\mathcal{T}$ then this BME score is equal to the PD value of X (on $\mathcal{T}$); while if d is additive on some other resolved tree $\mathcal{T}'$, then the BME score of $\mathcal{T}$ can be shown to exceed the PD value of set X (on $\mathcal{T}'$) [11]. The balanced minimum evolution method seeks the phylogenetic tree that minimizes the associated BME score. There is a close relationship between this method and Neighbor-Joining, which can be viewed as a locally optimal method for constructing a BME tree—for details see [12], [17].

6.5.1 *Tree reconstruction from PD-values over an abelian group*

So far we have regarded the lengths of the edges of a tree as being some positive real number. However, the concept of phylogenetic diversity is well-defined when

edge-weights are chosen from any abelian group $\mathcal{G}$ (briefly, an 'abelian group' is any set on which an addition can be defined which is associative and commutative, and there is a zero element and every element has an additive inverse; for details see [27]). This is both mathematically useful and potentially useful in applications. For the mathematical justification, one can ask what properties of PD depend on properties of the real numbers (such as the fact that they are ordered) and how much is just 'algebraic'. Clearly the 'Neighbor-Joining' algorithm no longer applies since the concept of minimizing or maximizing does not apply for a general abelian group. Moreover, although algebraic relations like the 3–point condition (equation (6.6)) apply in general, other results such as the representation (equation (6.2)) no longer do, as we may not be able to divide by factors such as $d(v) - 1$. Regarding tree reconstruction from pairwise PD values, the presence of elements of order 2 in a group (i.e. non-zero elements x for which $x + x = 0$) means that the classic uniqueness result no longer applies For example, consider the tree in Fig. 6.5, and the group $\mathbb{Z}_2 = \{0, 1\}$ under addition mod 2. Suppose the non-zero element (1) of this group is assigned to each edge of the tree shown in Fig. 6.5. Then we have $PD(\{x, y\}) = 0$ for any two elements x, y of the leaf set X of this tree. Moreover there exists more than one phylogenetic tree having this shape (in fact 15 such trees) so clearly PD values on pairs of elements of X are not sufficient to uniquely specify the underlying tree, in contrast to the case where the edges have real values.

It turns out, however, that if $\mathcal{G}$ has no elements of order 2 then the classic uniqueness (and existence) results for tree representations for pairwise PD values

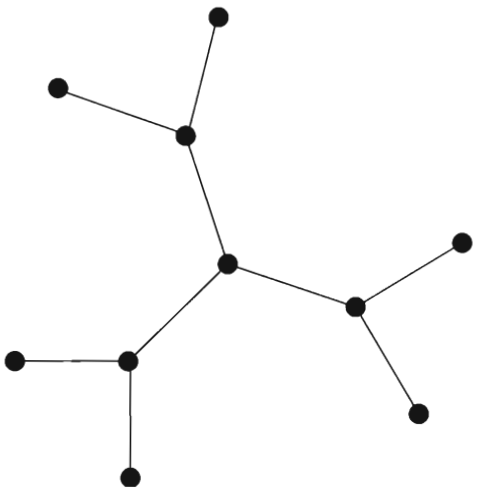


FIG. 6.5. Any leaf labelling of this tree gives $PD(\{x, y\}) = 0$ for all x, y when the element $1 \in \mathbb{Z}_2$ is assigned to each edge.

carry through to the abelian group setting. In the more general case where $\mathcal{G}$ may have elements of order 2, the uniqueness of a tree representation can be recovered, provided that one considers both pairwise and triplewise PD values [13].

More precisely the following result (from [13]) holds.

Theorem 6.6 *Let $\mathcal{T}$ be a phylogenetic X-tree, $\mathcal{G}$ an abelian group, and λ a function that assigns a non-zero element of $\mathcal{G}$ to each edge of $\mathcal{T}$. Then $\mathcal{T}$ is determined up to isomorphism (and can be reconstructed by an algorithm that runs in polynomial time in $|X|$) by the map that associates each pair and triple of elements of X with its associated $\mathcal{G}$-valued PD score.*

The existence question ('when can pairwise and triple-wise PD values be represented by a tree with edge weights drawn from an abelian group?') has also been settled—it involves the three-point condition (equation (6.6)), two four-point conditions, and a five-point condition. This last five-point condition is not required when $\mathcal{G}$ is the group of real numbers under addition, or indeed an abelian group without elements of order 2, but in general it is necessary (for details see [13]).

We end this section by outlining a situation in molecular biology where such group-based valuations arise naturally (the parity of gene orders provides another, but we will not describe this in detail here).

Consider DNA sequences of length k that have been re-coded as binary sequences (for example, by associating with each of the four bases its purine or pyrimidine class). Any two such binary sequences $(w_1, \ldots, w_k), (z_1, \ldots, z_k)$ define a $0-1$ sequence $g = (g_1, \ldots, g_k)$ of length k by setting $g_i = 0$ precisely if $w_i = z_i$, otherwise $g_i = 1$. We may regard g as an element of the abelian 2-group $\mathbb{Z}_2^k$. Now consider an evolutionary tree, where at each vertex there is some purine–pyrimidine sequence (carried by the ancestral taxon at that place in the tree). Assign to each edge the group element associated to its endpoints by the process just described. Then for any two leaves x, y the value $PD(\{x, y\})$ can be computed just from the sequences at x, y (without knowing the tree or the states assigned to other vertices)—it is simply the group element associated to the difference (or, equivalently, the sum) of the sequences at x and y. However, the value of $PD(\{x, y, z\})$ is not uniquely determined by just the sequences at x, y, and z (were this the case, then reconstructing phylogenetic trees from binary sequences would be essentially trivial). Determining $PD(\{x, y, z\})$ is equivalent to determining the sequence that was present at the *median vertex* in the tree connecting leaves x, y, z. This has a curious consequence—if one can reconstruct the ancestral sequence (of the median vertex) for any three binary sequences, then one can reconstruct the underlying tree. One might attempt to estimate this ancestral sequence as the (component-wise) median of the sequences at x, y, z but it turns out that in general the resulting PD values do not have a representation on any tree—indeed the condition for the existence of such a representation is that the splits induced by the sites of the binary sequences are compatible [13]. In practice, biological data would rarely be expected to fulfil this compatibility condition. Thus, more sophisticated approaches to estimate the ancestral sequence

at a median vertex (based on models of sequence evolution, and guided by the necessary three-, four- and five-point conditions) would need to be developed before such an approach to tree reconstruction could be applied for analysing DNA sequence data.

6.6 Concluding comments

In this chapter we have investigated several applications of phylogenetic diversity: biodiversity conservation, expected patterns in biodiversity loss, and phylogenetic tree construction. This wide range of applications poses some interesting mathematical problems and provides useful approaches for managing and exploring biodiversity and tree construction.

The Noah's Ark Problem (NAP) discussed here has been applied to both conservation and genomic sequencing problems. No efficient (polynomial time) algorithm for solving the general NAP is known to the authors, and a simple exhaustive search may need to consider a large proportion of the 2^n possible subsets of taxa (this is not feasible for a problem consisting of more than a few dozen taxa). As discussed, algorithms for efficiently computing solutions to several restricted variations of the NAP exist, but some suggestions have been made in the literature that the NAP is too simplistic and needs to incorporate more realistic aspects of the problem. These extensions will further complicate the problem of finding optimal solutions.

We have also illustrated the importance of the time scale of conservation management. The magnitude assigned to the survival probabilities of the taxa determines what management time scale is being considered. For non-trivial trees, the optimal solution to the NAP is sensitive to the time scale that has been selected; selecting an inappropriate time scale may result in an inappropriate prioritization of taxa to conserve.

Investigating the expected losses in PD as taxa become extinct, is a useful approach for quantifying future expected losses in biodiversity. Here we have shown that as taxa randomly become extinct, each new extinction is expected to cause a greater loss in biodiversity, though the rate of biodiversity loss with time exhibits a different behaviour. Further work using more realistic models of extinction could provide additional insight into the loss of biodiversity. It may be particularly relevant to consider survival probabilities (the a_j values) from a skewed distribution or correlated with the distance between taxa in the phylogenetic tree (Arne Mooers, pers. comm.).

Furthermore we have considered how PD may provide a useful tool for refining tree reconstruction by using m-way comparisons of taxa. For $m = 2$ this has been well studied, and is generally referred to as 'distance-based' approaches to tree reconstruction, however many results and methods (such as Neighbor-Joining) extend naturally to larger values of m.

A final generalization is to allow the branch lengths to take values in any abelian group. The message seems to be that for groups without elements of order 2, tree reconstruction behaves just like the familiar group of real numbers

(though some care is needed as concepts involving order and minimization no longer apply, so methods like Neighbor-Joining are problematic). For groups with elements of order 2, the mathematical analysis is slightly more complicated, but still tractable.

Acknowledgements

We thank Arne Mooers, Olivier Gascuel, and an anonymous referee for some helpful comments, and the *New Zealand Marsden Fund* and the *Allan Wilson Centre for Molecular Ecology and Evolution* for supporting this research.

References

[1] Altschul, S. F. and Lipman, D. J. (1990). Equal animals. *Nature*, **348** (6301), 493–494.

[2] Barker, G. M. (2002). Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the Linnean Society*, **76**, 165–194.

[3] Bertelli, S. and Giannini, N. P. (2005). A phylogeny of extant penguins (Aves: Spenisciformes) combining morphology and mitochondrial sequences. *Cladistics*, **21**, 209–239.

[4] Bunnell, F. L. and Huggard, D. J. (1999). Biodiversity across spatial and temporal scales: problems and opportunities. *Forest Ecology and Management*, **115**, 113–126.

[5] Camm, J. D., Norman, S. K., Polasky, S., and Solow, A. R. (2006). Nature reserve site selection to maximize expected species covered. *Operations Research*, **50**(6), 946–955.

[6] Clarke, K. R. and Warwick, R. M. (1998). A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology*, **35**, 523 531.

[7] Crozier, R. H. (1992). Genetic diversity and the agony of choice. *Biological Conservation*, **61**, 11–15.

[8] Crozier, R H (1997). Preserving the information content of species: Genetic diversity, phylogeny, and conservation worth. *Annual Review of Ecology and Systematics*, **28**, 243–268.

[9] Crozier, R. H., Agapow, P., and Dunnett, L. J. (2006). Conceptual issues in phylogeny and conservation: a reply to Faith and Baker. *Evolutionary Bioinformatics Online*, **2**, 197–199.

[10] Crozier, R. H., Dunnett, L. J., and Agapow, P. M. (2005). Phylogenetic biodiversity assessment based on systematic nomenclature. *Evolutionary Bioinformatics Online*, **1**, 11–36.

[11] Desper, R. and Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*, **21**(3), 587–598.

[12] Desper, R. and Gascuel, O. (2005). The minimum evolution distance-based approach to phylogenetic inference. In *Mathematics of Evolution and Phylogeny* (ed. O. Gascuel). Oxford University Press, New York.

[13] Dress, A. and Steel, M. (2006). Phylogenetic diversity over an abelian group. *Annals of Combinatorics*, In Press.

[14] Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61**, 1–10.

[15] Faith, D. P. and Baker, A. M. (2006). Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evolutionary Bioinformatics Online*, **2**, 70–77.

[16] Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability.* W. H. Freemand and Company, San Francisco.

[17] Gascuel, O. and Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, **23**(11), 1997–2000.

[18] Gaston, K. J. (1996). Species richness: measure and measurement. In *Biodiversity: A Biology of Numbers and Difference* (ed. K. Gaston), pp. 77–113. Blackwell Science, Cambridge.

[19] Giannini, N. P. and Bertelli, S. (2004, April). Phylogeny of extant penguins based on integumentary and breeding characters. *The Auk*, **121**(2), 422–434.

[20] Haake, C., Kashiwada, A., and Su, F. E. (2005, March). The shapley value of phylogenetic trees. *IMW Working Paper #363* (363).

[21] Hartmann, K. and Steel, M. (2006). Maximizing phylogenetic diversity in biodiversity conservation: Greedy solutions to the Noah's Ark Problem. *Systematic Biology*, **55**(4), 644–651.

[22] IUCN (2004). 2004 IUCN Red list of threatened species. `http://www.iucnredlist.org`.

[23] Korte, B., Lovász, L., and Schrader, R. (1991). *Greedoids, Algorithms and Combinatorics.* Springer-Verlag Berlin.

[24] Levy, D., Yoshida, R., and Pachter, L. (2006). Neighbor joining with phylogenetic diversity estimates. *Molecular Biology and Evolution*, **23**(3), 491–498.

[25] Lewis, C. A., Lester, N. P., Bradshaw, A. D., Fitzgibbon, J. E., Fuller, K., Hakanson, L., and Richards, C. (1996). Considerations of scale in habitat conservation and restoration. *Canadian Journal of Fisheries and Aquatic Sciences*, **53**(Suppl. 1), 440–445.

[26] Lewis, L. A. and Lewis, P. O. (2005). Unearthing the molecular phylodiversity of desert soil green algae (Chlorophyta). *Systematic Biology*, **54**(6), 936–947.

[27] Maclane, S. and Birkoff, G. (1979). *Algebra* (second edn). Macmillan, New York.

[28] Minh, B. Q., Klaere, S., and von Haeseler, A. (2006). Phylogenetic diversity within seconds. *Systematic Biology*, **55**(5), 769-773.

[29] Mooers, A. Ø., Heard, S. B., and Chrostowski, E. (2005). Evolutionary heritage as a metric for conservation. In *Phylogeny and Conservation* (ed. A. Purvis, T. Brooks, and J. Gittleman), pp. 120–138. Cambridge University Press, New York.

[30] Nee, S., and May, R. M. (1997). Extinction and the loss of evolutionary history. *Science*, **278**(5338), 692–694.

[31] Norton, B. G. (1987). *Why Preserve Natural Variety?* Princeton University Press, Princeton.

[32] Pachter, L. and Speyer, D. (2004). Reconstructing trees from subtree weights. *Applied Mathematics Letters*, **17**(6), 615–621.

[33] Pardi, F. and Goldman, N. (2007). Resource aware taxon selection for maximising phylogenetic diversity. *Systematic Biology*, In Press.

[34] Pardi, F. and Goldman, N. (2005). Species choice for comparative genomics: no need for cooperation. *PLoS Genetics*, **1**(6), 71.

[35] Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, **51**, 41–47.

[36] Pavoine, S., Ollier, S., and Dufour, A. (2005). Is the originality of a species measurable? *Ecology Letters*, **8**, 579–586.

[37] Pullin, A. S. (2002). *Conservation Biology*. Cambridge University Press, New York.

[38] Redding, D. W., and Mooers, A. Ø. (2006). Incorporating evolutionary measures into conservation prioritization. *Conservation Biology*, In Press.

[39] Reist-Marti, S., Abdulai, A., and Simianer, H. (2006). Optimum allocation of conservation funds and choice of conservation programs for a set of African cattle breeds. *Genetics Selection Evolution*, **38**, 99–126.

[40] Rodrigues, A. S. L., Brooks, T. M., and Gaston, K. J. (2005). Integrating phylogenetic diversity in the selection of priority areas for conservation: does it make a difference? In *Phylogeny and Conservation* (ed. A. Purivs, J. L. Gittleman, and T. Brooks), Number 8 in Conservation Biology, Chapter 5, pp. 101–119. Cambridge University Press, New York.

[41] Sechrest, W., Brooks, T. M., da Fonseca, G. A. B., Konstant, W. R., Mittermeier, R. A., Purvis, A., Rylands, A. B., and Gittleman, J. L. (2002). Hotspots and the conservation of evolutionary history. *Proceedings of the National Academy of Sciences*, **99**(4), 2067–2071.

[42] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press, New York.

[43] Semple, C. and Steel, M. (2004). Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, **32**(4), 669–680.

[44] Simianer, H., Marti, S. B., Gibson, J., Hanotte, O., and Rege, J. E. O. (2003). An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds. *Ecological Economics*, **45**, 377–392.

[45] Soutullo, A., Dodsworth, S., Heard, S. B., and Mooers, A. Ø. (2005). Distribution and correlates of carnivore phylogenetic diversity across the Americas. *Animal Conservation*, **8**(3), 249–258.

[46] Steel, M. (2005). Phylogenetic diversity and the greedy algorithm. *Systematic Biology*, **54**(4), 527–529.

[47] Steel, M. (2006). Tools to construct and study big trees: A mathematical perspective. In *Reconstructing the Tree of Life: Taxonomy and Systematics of Species Rich Taxa* (ed. T. R. Hodkinson and J. A. Parnell). CRC Press.

[48] Steel, M. A., Penny, D., and Hendy, M. D. (1988). Loss of information in genetic distance. *Nature*, **336**(6195), 118.

[49] Sumner, J. G., and Jarvis, P. D. (2005). Entanglement invariants and phylogenetic branching. *Journal of Mathematical Biology*, **51**(1), 18–36.

[50] van der Heide, C. M., van den Bergh, Jeroen C. J. M., and van Ierland, E. C. (2005). Extending Weitzman's economic ranking of biodiversity protection: combining ecological and genetic considerations. *Ecological Economics*, **55**(2), 218–223.

[51] Vane-Wright, R. I., Humphries, C. J., and Williams, P. H. (1991). What to protect? - Systematics and the agony of choice. *Biological Conservation*, **55**, 235–254.

[52] Weitzman, M. L. (1998). The Noah's Ark Problem. *Econometrica*, **66**(6), 1279–1298.

[53] Wilson, K. A., McBride, M. F., Bode, M., and Possingham, H. (2006). Prioritizing global conservation efforts. *Nature*, **440**, 337–340.

[54] Zaretskii, K. A. (1965). Constructing trees from the set of distances between pendant vertices. *Uspehi Matematiceskih Nauk*, **20**, 90–92.