Syst. Biol. 55(4):644–651, 2006

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150600873876

Maximizing Phylogenetic Diversity in Biodiversity Conservation: Greedy Solutions to the Noah's Ark Problem

KLAAS HARTMANN AND MIKE STEEL

Allan Wilson Centre for Molecular Ecology and Evolution, Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand; E-mail: m.steel@math.canterbury.ac.nz (M.S.)

Abstract.—The Noah's Ark Problem (NAP) is a comprehensive cost-effectiveness methodology for biodiversity conservation that was introduced by Weitzman (1998) and utilizes the phylogenetic tree containing the taxa of interest to assess biodiversity. Given a set of taxa, each of which has a particular survival probability that can be increased at some cost, the NAP seeks to allocate limited funds to conserving these taxa so that the future expected biodiversity is maximized. Finding optimal solutions using this framework is a computationally difficult problem to which a simple and efficient "greedy" algorithm has been proposed in the literature and applied to conservation problems. We show that, although algorithms of this type cannot produce optimal solutions for the general NAP, there are two restricted scenarios of the NAP for which a greedy algorithm is guaranteed to produce optimal solutions. The first scenario requires the taxa to have equal conservation cost; the second scenario requires an ultrametric tree. The NAP assumes a linear relationship between the funding allocated to conservation of a taxon and the increased survival probability of that taxon. This relationship is briefly investigated and one variation is suggested that can also be solved using a greedy algorithm. [Biodiversity conservation; greedy algorithm; Noah's Ark Problem; phylogenetic diversity.]

Biodiversity conservation requires a methodology for prioritizing the taxa to conserve, given limited resources. Many conservation approaches have simply aimed to conserve as many taxa as possible (Gaston 1996); however, a more appropriate method should take taxon distinctiveness into account (for review, see Crozier, 1997) and aim to minimize the future loss of biodiversity. Various simple indices have been developed that give an indication of the distinctiveness of a taxon or of its importance to the future conservation of biodiversity. These indices are based on the phylogenetic tree containing the taxa of interest, notable examples include Vane-Wright et al. (1991), May (1990), Faith (1992), Haake et al. (2005), and Redding and Mooers (2006). A limitation of most such indices is that they do not consider the differential costs involved in conserving different taxa or the different survival probabilities of different taxa. However, Witting and Loeschcke (1995) (see also, Witting et al., 2000) linked the phylogenetic diversity (*PD*) measure (Faith 1992) to extinction probabilities to obtain a method for minimizing the future loss of biodiversity.

Weitzman (1998) proposed the "Noah's Ark Problem" (NAP), a framework based on *PD* that incorporates costs and probabilities and has seen some practical application including conservation of cattle breeds (Simianer 2003; Reist-Marti et al., 2006). In the NAP each taxon has a survival probability that can be increased at some cost. The objective is to allocate a limited budget to the taxa such that the future expected biodiversity (as obtained from the phylogenetic tree) is maximized. Unfortunately, obtaining this optimal budget allocation is a complex problem and it may be necessary to consider a large proportion of the possible subsets of the N taxa that can be conserved. The number of such subsets grows at rate 2^N, consequently for problems involving more than a few dozen taxa, it is not computationally feasible to consider all subsets and an efficient algorithm is required for obtaining optimal solutions to the NAP.

Steel (2005) considered a simplified version of the NAP in which taxa only survive if they are conserved and all taxa cost the same to conserve. Steel (2005) showed that optimal solutions to this problem can be produced using a simple and efficient greedy algorithm. In this paper we investigate two more realistic variations of the NAP that allow for variable conservation costs and uncertain survival of the taxa. These variations are also shown to be solvable in polynomial time using a greedy algorithm.

Suggestions have been made in the literature that any NAP for which the associated tree satisfies a molecular clock can be solved using a greedy algorithm (Simianer, 2003). Several aspects of the NAP that prevent the greedy algorithm from producing optimal solutions in all cases are examined. These examples (Figures 2 and 5) illustrate that a greedy algorithm is not, in general, guaranteed to produce optimal solutions.

Phylogenetic Diversity

The NAP uses phylogenetic diversity (*PD*; Faith, 1992) as a measure of biodiversity. *PD* has been used in a wide variety of applications including biodiversity conservation (e.g., Crozier et al., 2005; Lewis and Lewis, 2005; Mooers et al., 2005; Soutullo et al., 2005; and Faith and Williams, 2006) and prioritizing taxa for genomic sequencing (Pardi and Goldman, 2005). *PD* is calculated from the phylogenetic tree $\mathcal{T}$, the leaves of which correspond to the set of taxa, X, under study. For a subset Y of X the *PD* is the sum of the branch lengths of the phylogenetic tree containing taxa in Y and the root, an example is given in Figure 1. Note that a variant of *PD* has been used elsewhere where the root is not necessarily included. The standard usage appears to include the root (Faith, 1992) (see Faith and Baker, 2006, in response to Crozier et al., 2005, for further discussion).

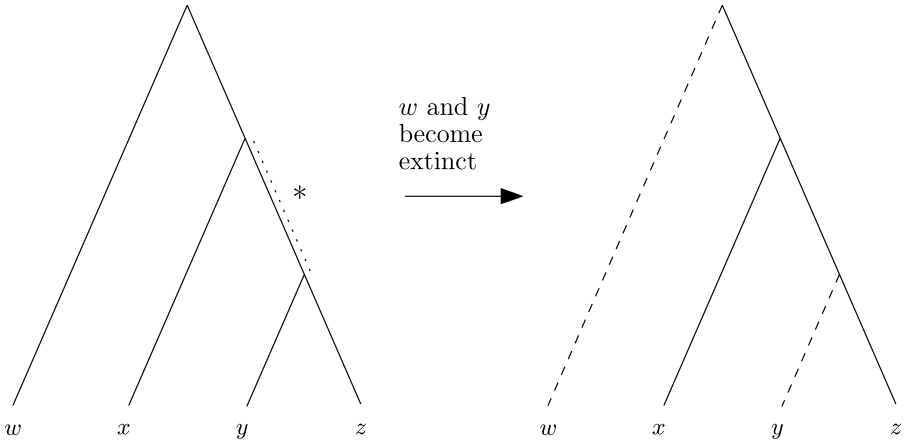


FIGURE 1. When taxa w and y in the tree on the left become extinct, the edges that are considered to have been preserved are indicated by the solid lines. The phylogenetic diversity (*PD*) is calculated as the sum of the preserved edge lengths. The edge indicated by * is considered further in the text.

If each taxon remains extant until some given future time with probability a_j, then it is possible to calculate the expected *PD* at that time. A particular branch length is included in the *PD* score if at least one of the children of that edge remains extant. For example, the edge indicated by an * in Figure 1 will be preserved as long as one of its children (taxa y or z) remains extant. If these taxa have a survival probability of 0.9, the probability that at least one will remain extant (and the edge * is preserved) is simply $1-(1-0.9)^2 = .99$. Denoting the children of a particular edge, i, by C_i the expected *PD* can in general be expressed as:

$$\mathbb{E}(PD) := \sum_i \lambda_i \left(1 - \prod_{j \in C_i} (1 - a_j)\right), \qquad (1)$$

where λ_i is the length of edge i, and the summation is over all edges of $\mathcal{T}$. 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 paper 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 (recall that an *ultrametric tree* is one for which the distance from the root to any leaf is the same, as would occur for genetic distance under a molecular clock, or under an evolutionary time-scale).

The Noah's Ark Problem

A variation of the Noah's Ark Problem was described in Weitzman (1992); however, this used a measure of dissimilarity instead of *PD* (see Faith, 2003, for a discussion); the NAP as published in Weitzman (1998) finally combined *PD*, extinction probabilities, and conservation costs.

In the NAP framework each taxon, j, has some probability, a_j, of remaining extant; however, 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. Given a budgetary constraint, B, the problem is to find the set of taxa to conserve, S, that maximizes the future expected phylogenetic diversity, denoted by $\mathbb{E}(PD|S)$. The quantity $\mathbb{E}(PD|S)$ is calculated in a similar fashion to Equation 1, except that the survival probabilities of the conserved taxa (those taxa in S) need to be considered separately:

$$\begin{aligned}\mathbb{E}(PD|S) &= \sum_i \lambda_i\, p(i|S) \\ &= \sum_i \lambda_i \left(1 - \prod_{k \in C_i - S} (1 - a_k) \prod_{l \in C_i \cap S} (1 - b_l)\right), \qquad (2)\end{aligned}$$

where $p(i|S)$ denotes the probability that one of the taxa in C_i will remain extant given that the set of taxa S is being conserved.

The formulation of the NAP used throughout this paper is essentially equivalent to that given in Weitzman (1998) but is expressed differently for convenience:

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 constraint ensures that the cost of conserving the taxa in S does not exceed the budget (B).

The original formulation of the NAP included an additional term in the objective function that permitted each taxon to have an intrinsic value (utility) unrelated to its contribution to *PD* (e.g., the value of tourism for a species

of whale). This additional value has not been made explicit here, as it is easy to show that including such a value for a particular taxon is equivalent to adding it to the length of the pendant edge for that taxon.

The original formulation of the NAP also allowed taxa to be partially protected, so that resources could be spread more thinly across multiple taxa instead of conserving a smaller subset of taxa to the maximum extent possible. Weitzman (1998) assumed that if a taxon is partially conserved, the survival probability increase for that taxon is directly proportional to the proportion of the funding that taxon received. If q_j is spent on conserving taxon j ($0 \leq q_j \leq c_j$), the new survival probability, $g_j(q_j)$, for taxon j is:

$$g_j(q_j) = \frac{q_j}{c_j}(b_j - a_j) + a_j. \tag{3}$$

Weitzman (1998) showed that under this assumption, the solutions to the NAP are extreme—the optimal solution will always allocate the maximum amount ($q_j = c_j$) to a few taxa instead of partially conserving ($0 < q_j < c_j$) a greater number of taxa, with the possible exception of the last taxon conserved, which may only be partially conserved due to budgetary constraints. Consequently, the problem of deciding how much funding to allocate to the conservation of each taxon becomes a problem of deciding which taxa to conserve. Throughout this paper we will adopt the convention that the last taxon selected for conservation will be partially conserved such that the full conservation budget is utilized.

The benefit of Equation (3) is further demonstrable by considering a star tree (each taxon is directly descendant from the root) where the taxa may have different costs. If $a_j = 0$ and $b_j = 1$ for all taxa j and each taxon can be either fully conserved or not at all:

$$g_j(q_j) = \begin{cases} 0, & q_j < c_j; \\ 1, & q_j \geq c_j; \end{cases}$$

then the problem is equivalent to the "knapsack problem," which is well known to be NP-complete (Cormen, 2001). However, if $g_j(q_j)$ is given by Equation (3) instead, the problem is equivalent to the "fractional knapsack" problem, which is solvable by a greedy algorithm (Cormen et al., 2002).

Scenario 1: Constant Conservation Costs and Variable Survival Probabilities

Steel (2005) considered a variant of the NAP where the conservation cost is the same for all taxa ($c_j = c$) and taxa only survive if they are conserved ($a_j = 0$, $b_j = 1$). That paper established that all optimal solutions for this problem can be produced by a greedy algorithm that builds up a set by sequentially adding the taxon that produces the greatest increases in *PD*. In Hartmann and Steel (2006) this result was extended to allow non-zero survival probabilities in the absence of conservation ($a_j \neq 0$). Here we provide a further extension which permits all survival rates to be non-zero/non-unity.

Theorem 1. *For the Noah's Ark problem with equal conservation costs optimal solutions can be produced by a greedy algorithm if the following condition is met by the survival probabilities*:

$$\frac{1 - b_j}{1 - a_j} = \kappa, \tag{4}$$

for some constant κ *(with* $0 \leq \kappa \leq 1$*). The algorithm begins with an empty set* S *and sequentially adds the taxon,* j*, which maximizes* $\mathbb{E}(PD|S \cup j)$ *until* S *is at the maximum size permitted by the budgetary constraint.*

Note that, if conservation is completely efficient ($b_j = 1$), the survival probabilities in the absence of conservation (a_j) are free to vary; otherwise, this condition states that the extinction probability must be reduced by the same proportion for each taxon when it is conserved [$1 - b_j = \kappa(1 - a_j)$].

Proof. The proof proceeds in a similar fashion to Steel (2005) by establishing a strong exchange property: namely that for any two subsets, Y and Z, of X with $|Y| < |Z|$ there exists some taxon $z \in Z$ such that:

$$\mathbb{E}(PD|Z - \{z\}) - \mathbb{E}(PD|Z) + \mathbb{E}(PD|Y \cup \{z\}) - \mathbb{E}(PD|Y) \geq 0. \tag{5}$$

This means that for any two subsets of X, the larger subset contains some taxon (z) that would contribute more to the expected *PD* value of the smaller subset than it adds to that of the larger one.

Denote the set of edges on the path from z to the root by R, and notice that each of the expected *PD* terms in Equation (5) can be split into a sum over the edges in R, and a sum over the edges not in R. The significance of this observation is that the probability that edges not in R are spanned remains unchanged as z is removed from Z or added to Y. Denoting the left-hand side of Equation (5) by ΔPD we have:

$$\Delta PD = \sum_{i \in R} \lambda_i \Delta p(i) + \sum_{j \notin R} \lambda_j \Delta p(j),$$

where

$$\Delta p(i) := p(i|Z - \{z\}) - p(i|Z) + p(Y \cup \{z\}) - p(i|Y),$$

then for $j \notin R$ we have $\Delta p(j) = 0$ because the probability of an edge not in R being spanned is independent of the presence of taxon z, hence:

$$\Delta PD = \sum_{i \in R} \lambda_i \Delta p(i).$$

A sufficient condition for satisfying the strong exchange property (Equation (5)) is therefore that $\Delta p(i) \geq 0$ for each edge i on the path from taxon z to the root. The following results follow from the definition of $p(i|Z)$:

$$\begin{aligned} &p(i|Z - \{z\}) - p(i|Z) \\ &\quad = (a_z - b_z) \prod_{m \in C_i - Z} (1 - a_m) \prod_{l \in C_i \cap Z - \{z\}} (1 - b_l) \\ &p(i|Y \cup \{z\}) - p(i|Y) \\ &\quad = (b_z - a_z) \prod_{m \in C_i - Y - \{z\}} (1 - a_m) \prod_{l \in C_i \cap Y} (1 - b_l). \end{aligned}$$

Combining these gives an identity for $\Delta p(i)$ which can be further simplified using Equation (4):

$$\begin{aligned} \Delta p(i) &= \left[\prod_{m \in C_i - Z} (1 - a_m) \prod_{l \in C_i \cap Z - \{z\}} (1 - b_l) \right. \\ &\quad \left. - \prod_{m \in C_i - Y - \{z\}} (1 - a_m) \prod_{l \in C_i \cap Y} (1 - b_l) \right] (a_z - b_z) \\ &= \left(\kappa^{|C_i \cap Z| - 1} - \kappa^{|C_i \cap Y|}\right)(a_z - b_z) \prod_{m \in C_i} (1 - a_m) \end{aligned}$$

Noting that $a_z - b_z$ is negative, a sufficient condition for insuring that $\Delta p(i) \geq 0$ is $\kappa^{|C_i \cap Z| - 1} - \kappa^{|C_i \cap Y|} \leq 0$, which (because $0 \leq \kappa \leq 1$) is equivalent to

$$|C_i \cap Y| \leq |C_i \cap Z| - 1. \tag{6}$$

This condition simply states that the number of elements in Y that span edge i is strictly less than the number of elements in Z that span that edge. Next we show that for any two sets Y and Z with $|Y| < |Z|$, it is possible to find a taxon z for which this last property holds for each edge i on the path from z to the root.

Starting at the root, one of the edges adjacent to the root must satisfy Equation (6) because $|Y| < |Z|$ (if Equation (6) were not satisfied this would imply $|Y| \geq |Z|$), call this edge m. Similarly, one of the edges below m must satisfy Equation (6) because we have $|C_m \cap Y| \leq |C_m \cap Z| - 1$, pick this edge, call it m, and continue this procedure until one arrives at an exterior edge. The condition $\Delta p(i) \geq 0$ is therefore met on every edge, from the taxon adjacent to this exterior edge through to the root, consequently the strong exchange property (Equation (5)) holds.

Let Y be an optimal solution if m taxa are to be conserved and Z an optimal solution if $m + 1$ taxa are to be conserved. Applying the strong exchange property (Equation (5)) to Y and Z shows the existence of a taxon z such that $Y \cup \{z\}$ is an optimal solution for $m + 1$ taxa and $Z - \{z\}$ is an optimal solution for m taxa.

Theorem 1 follows easily by standard arguments from "greedoid" theory (Korte et al., 1991). Specifically, the above observation shows that any solution for $m + 1$ taxa must be obtained from a solution for m taxa by adding a single taxon which maximizes the increase in $\mathbb{E}(PD|Y)$.

The Necessity of Equation (4)

Any problem for which the greedy algorithm is optimal must satisfy the substructure property (Cormen et al., 2002). This property states that an optimal solution, Y, of a given size must be contained within an optimal solution of each larger size. The condition imposed in the previous section (Equation (4)) ensures that the substructure property holds for the optimization problem.

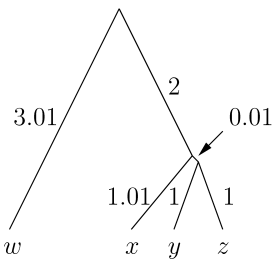
Here we provide a simple example to show that this substructure property (and thereby the greedy algorithm) can fail when the condition imposed by Equation (4) in Theorem 1 is violated.

In Figure 2 the optimal subset of size 1 is $\{x\}$. The additional contribution to $E(PD)$ made by the pendant edge of x when it is conserved is smaller than that from the pendant edges of y or z (were they to be conserved). The optimality of x is entirely due to its conservation ensuring that the interior edge of length 2 is spanned.

When two taxa are conserved, the probability increase that x provides for the interior edge of length 2 is reduced such that the smaller increase in this probability that y and z provide, coupled with the greater contribution from their pendant edges, makes x a less valuable taxon to conserve. The optimal subset of size 2 is therefore $\{y, z\}$ (see Figure 2), the substructure property is violated (which was possible as the condition in the previous section [Equation (4)] was not satisfied) and the greedy algorithm cannot produce the optimal solution.

Nonlinear Conservation Expenditure and Taxon Survival Relationship

Recall that the expenditure-survival relationship $g_j(q_j)$ gives the probability that a taxon, j, will remain extant given that q_j is spent on its conservation. Scenario 1 and Scenario 2 (in the following section) assume a linear relationship for $g_j(q_j)$ (Equation (3)). This linear relationship ensures that solutions are extreme—all taxa with one possible exception are fully conserved or not at all—which in turn simplifies the NAP problem from one of deciding the amount to spend on the conservation of



Subset (S)	$\mathbb{E}(PD\|S) - \mathbb{E}(PD\|\emptyset)$
$\{w\}$	0.602
$\{x\}$	1.068
$\{y\}$,$\{z\}$	1.055
$\{w, y\}$,$\{w, z\}$	1.657
$\{w, x\}$	1.670
$\{y, z\}$	1.745
$\{x, y\}$, $\{x, z\}$	1.672

FIGURE 2. A NAP that does not satisfy condition 4 and violates the substructure property. The optimal subset of size 1 is $\{x\}$ and the optimal subset of size 2 is $\{y, z\}$. Parameter values are $a_w = 0.6$; $a_x = 0.5$; $a_y = a_z = 0.25$; $b_w = 0.8$; $b_x = 1$; and $b_y = b_z = .85$.

each taxon to one of selecting the optimal set of taxa to conserve.

Simianer et al. (2003) questioned the validity of the linear relationship and applied the NAP using various alternatives to Equation (3). Further examples of different relationships can be found in Johst et al. (2002) and Lamberson et al. (1992).

For convenience, problems with $g_j(q_j)$ not of the type given in Equation (3) will be referred to as Generalized Noah's Ark Problems (g-NAPs). The relationships $g_j(q_j)$ within a g-NAP are generally not parameterized by a_j, b_j, and c_j and cannot be assumed to have extreme solutions. However, as we will show, there is one family of g-NAPs that can be solved using a greedy algorithm, namely g-NAPs, where $g_j(q_j)$ has the form:

$$g_j(q_j) = 1 - k^{q_j}(1 - a_j) \text{ with } \quad 0 \leq k \leq 1,$$

can be transformed to a NAP (with $g_j(q_j)$ as in Equation (3)) using the method detailed in Appendix 1; the resulting NAP is of the type described in Scenario 1. Consequently such problems can be solved using a greedy algorithm.

This formulation of $g_j(q_j)$ corresponds to the situation where each budgetary unit allocated to conserving a taxon produces progressively smaller increases in that taxon's survival probability as dictated by the above equation. Note that survival of a taxon cannot be guaranteed regardless of the funding allocated to its conservation (unless of course $a_j = 1$).

Other g-NAPs that satisfy certain conditions (discussed in Appendix 1) can be transformed to NAPs. The resulting NAPs will generally not fall into Scenario 1 and may therefore violate the substructure property, hence they may not be solvable using a greedy algorithm.

Scenario 2: Variable Conservation Costs and an Ultrametric Tree

In this section a variation of the NAP is considered that allows variable conservations costs and for which the greedy algorithm can produce an optimal solution (W).

Denote the expected contribution a particular taxon, j, makes to the expected PD of a set of taxa, W, by $\sigma_W(j)$. That is, if j is in W, $\sigma_W(j)$ is the PD that W would lose if j were removed, if j is not in W it is the PD that W can gain from the addition of j:

$$\sigma_W(j) := \mathbb{E}(PD|W \cup \{j\}) - \mathbb{E}(PD|W - \{j\}).$$

The cost-benefit of adding a taxon to a subset is given by $r_W(j) = \sigma_W(j)/c(j)$, this is the contribution j makes to the PD per unit of cost. The overall cost benefit of a particular subset of taxa W is $R_W = \mathbb{E}(PD|W)/\sum_{j \in W} c(j)$, and optimal solutions to the NAP will maximize R_W subject to the total cost equaling the conservation budget (B).

Theorem 2. *A greedy algorithm produces optimal solutions for any Noah's Ark Problem with variable conservation costs provided the tree is ultrametric and conservation increases the survival probability of each taxon from certain extinction ($a_j = 0$) to certain survival ($b_j = 1$).*

The greedy algorithm begins with $W = \emptyset$ and continues to add the taxon with the highest value of $r_W(i)$ to W until the cost of conserving the taxa in W exceeds the budget. The last taxon added should be partially conserved to bring the total cost to the budget.

This theorem is a variation of that stated, without reference or proof, in Weitzman (1992: 374) and Weitzman (1995: 31). The difference between the proposed algorithms is that the greedy algorithm presented here builds up a set of taxa to conserve by adding one taxon at a time, whereas that proposed by Weitzman begins with the full set of taxa and removes one taxon at a time. The requirement in Weitzman (1992) that the dissimilarity measure be ultrametric and the requirement in Weitzman (1995) of a bead model of evolutionary branching are both equivalent to requiring the tree to be ultrametric. Weitzman's theorem claims that the greedy algorithm will produce optimal results for an ultrametric tree and it allows for intrinsic values of the conserved taxa (as discussed previously). However, it is the modified tree where the intrinsic values of the taxa have been added to the pendant edges that must be ultrametric.

Proof. Theorem 2 cannot be proven in the same manner as Theorem 1 because the strong exchange property (Equation (5)) does not hold (it is a straightforward matter to construct a counterexample). Instead, for this scenario we establish two claims: (i) all subsets not produced by the greedy algorithm are suboptimal and (ii) all subsets produced by the greedy algorithm are optimal.

Claim (i). Suppose that W is an optimal subset that can not be produced by the greedy algorithm. Consider constructing W by beginning with an empty set and adding the elements in W one at a time such that a greedy choice is made whenever possible. Because W cannot be produced by a greedy algorithm, there will be some point in this sequence where a taxon, h, is added instead of a greedy choice, denote the subset to which h is added by Y ($Y \subset W$), and a taxon that the greedy algorithm would have added by g ($g \in W$, $g \notin Y$).

Consider the taxon in W that is the closest to g, without loss of generality the situation is as depicted in Figure 3. Denote this taxon by j (this taxon may not be unique, in this case the choice of j is arbitrary). It is necessary to consider two cases: $j \in W - Y$ and $j \notin W - Y$.

If $j \in W - Y$, g was a greedy choice at a time where j could have been added to the subset Y and because the greedy choice was not made we have:

$$r_Y(g) > r_Y(j),$$

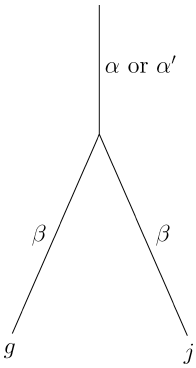


FIGURE 3. The general situation when two taxa, g and j, that share a common edge not in $\mathcal{T}|Y$ are added to $\mathcal{T}|W$. The tree has been assumed to be ultrametric. The root of the depicted tree corresponds to an interior node of $\mathcal{T}|Y$ or $\mathcal{T}|W$ when the length of the root edge is α or α', respectively.

that is,

$$\frac{\alpha+\gamma}{c_g} > \frac{\alpha+\gamma}{c_j} \text{(using the branch lengths in Figure 3)}$$
$$c_g < c_j. \tag{7}$$

The cost benefits of g and j relative to the final subset (W) are:

$$r_W(g) = (\alpha'+\gamma)/c_g,$$
$$r_W(j) = (\alpha'+\gamma)/c_j.$$

From Equation (7) we have $c_g < c_j$, hence $r_W(g) > r_W(j)$. The cost benefit of g exceeds that of j; diverting some funding from taxon j to g will increase the overall cost benefit, hence W is not an optimal subset.

If $j \notin W - Y$ there is no taxon in $W - Y$ that can reduce the cost benefit of g, hence the cost benefit of g still exceeds that of h and diverting funding from h to g will again increase the overall cost benefit, hence W is not an optimal subset. Hence, all optimal solutions must be produced by a greedy algorithm.

Claim (ii). Because an optimal solution exists (but may not be unique) at least one solution produced by the greedy algorithm must be optimal. To show that all solutions are in fact optimal it suffices to examine what happens when the greedy algorithm has to select from several greedy choices to add to a subset Y. Consider the case where there are two taxa, j and k with equal cost benefit. This can occur in two ways as depicted in Figure 4.

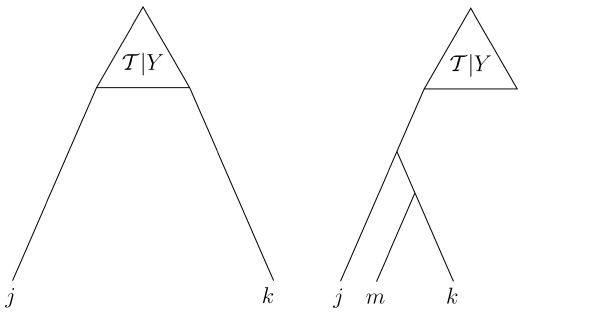


FIGURE 4. The two ways in which taxa with the same cost benefit may attach to an existing tree, $\mathcal{T}|Y$. Note that in both cases there may be any number of other taxa not in Y that attach to the edges depicted (such as the taxon m).

Case 1

The taxa with equal cost benefit attach to different internal nodes of $\mathcal{T}|Y$. In this case, addition of either taxa does not effect the cost benefit of the other taxon; regardless of which taxon is conserved first the other will be conserved next at the same cost benefit.

Case 2

The taxa with equal cost benefit attach to the same internal node of $\mathcal{T}|Y$. This situation is more complex, addition of the first taxon reduces the cost benefit of the second taxon consequently other taxa may have a higher cost benefit and be conserved before the second taxon.

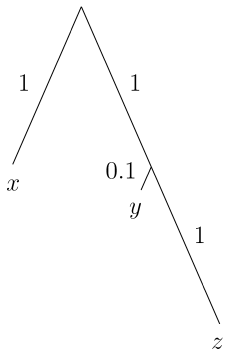
As j and k have the same cost benefit, the fact that the tree is ultrametric dictates that j and k have the same cost. It is therefore apparent that both the remaining budget and the cost benefit of the unconserved taxa are independent of which of j and k is conserved first. Only the cost benefits of those taxa that are incident with the pendant edge of j or k in $\mathcal{T}|Y \cup \{j,k\}$ (for example taxon m in Figure 4) are dependent on which of j and k is first conserved. However, from the same argument used to produce Equation (7), all of these taxa will have a higher cost than j and k, subsequently they will not be conserved until both j and k have been conserved (at which point it becomes irrelevant which of these taxa was conserved first).

The extension to more than two taxa with the same cost benefit, possibly with combinations of these two scenarios is straightforward.

Beyond Ultrametric Trees

When applied to a tree that is not ultrametric, the greedy algorithm is no longer guaranteed to provide the optimal solution. In particular when new taxa are added by the greedy algorithm, it is possible for taxa that have been added previously to have their cost benefit reduced below that of some taxon not selected thus far—this problem may not exhibit the substructure property. This is illustrated in Figure 5. The optimal subset of size 1 is $\{b\}$ with a cost benefit of 1, whereas the optimal subset of size two is $\{a, c\}$ with a cost benefit of 12/17.

Note that this problem is equivalent to a problem where the pendant edges of y and z have zero length and x, y, and z have intrinsic values of 0, 0.1, and 1, respectively. The resulting tree is ultrametric and thus by the theorem proposed in Weitzman (1992: 374) and Weitzman (1995: 31) should be solveable by their greedy algorithm. However, because the optimal solutions do



Subset (S)	Overall Cost Benefit (W_S)
$\{x\}$	0.5
$\{z\}$	0.89
$\{y\}$	1.1
$\{y, z\}$	0.65
$\{x, y\}$	0.70
$\{x, z\}$	0.71

FIGURE 5. A tree that is not ultrametric can lead to a violation of the substructure property. The optimal subset of size 1 is $\{y\}$ and the optimal subset of size 2 is $\{x, z\}$. Parameter values are $c_x = 2$, $c_y = 1$, $c_z = 2\frac{1}{4}$, and edge lengths as indicated.

not satisfy the substructure property, they cannot be produced by any greedy algorithm. If intrinsic values are being considered, the tree formed when these values are added to the pendant edges must be ultrametric for a greedy algorithm to produce optimal solutions.

REMARKS

Simple greedy algorithms were outlined for two special cases of the Noah's Ark Problem (NAP). These special cases are more realistic than that considered by Steel (2005) for which it is known that a greedy algorithm exists. Using these algorithms optimal solutions for practical problems that fall within these scenarios can be computed efficiently.

Simianer (2003: 384) has suggested (without proof) that a greedy algorithm will produce optimal solutions for a family of problems equivalent to the Generalized Noah's Ark Problem (g-NAP) described here, provided the tree satisfies a molecular clock. This family of problems includes the NAP proposed by Weitzman (1998) for which we have illustrated several cases where a greedy algorithm cannot produce optimal solutions (Figures 2 and 5). Hence we have shown that greedy algorithms are not, in general, guaranteed to produce optimal solutions for NAPs or g-NAPs. Caution is advised when applying a greedy algorithm to a problem not of the types described in Scenarios 1 and 2—the solutions produced may not be optimal.

Reist-Marti et al. (2006) describe a two-step algorithm for solving g-NAPs; they note that this algorithm is not guaranteed to produce the optimal solution. Algorithms such as this may prove useful, particularly for more complicated variations of the NAP. It would also be of interest to determine how close the solutions produced by such algorithms are to the global optimal.

Further extensions to the NAP to improve realism have been suggested in the literature. Simianer et al. (2003) suggested using more realistic relationships between the expenditure on conserving a taxon and that taxon's survival probability. A family of relationships that can be solved using the greedy algorithm has been presented here, it is expected that for most other relationships no guarantee can be made that the greedy algorithm will produce optimal solutions. van der Heide et al. (2005) suggested that interdependent survival probabilities derived from ecologically relationships should also be considered; if, for example, a species of prey becomes extinct, then the predators of that species may be more likely to become extinct. This has not been considered here but is worthy of further investigation.

Finally, a variation on the NAP is to select a subset S of taxa to maximize the probability $\mathbb{P}(PD|S \geq l)$ that the evolutionary heritage exceeds a value l given that the species in S are conserved and subject to the usual budget constraint $\sum_{j \in S} c_j \leq B$. This problem may be as relevant for biodiversity conservation as the standard NAP, though it is not clear if it would be as mathematically tractable to analyze.

ACKNOWLEDGMENTS

We thank Arne Mooers, Dan Faith, Fabio Pardi, and an anonymous referee for some helpful comments and the Allan Wilson Centre for helping fund this research.

REFERENCES

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2002. Introduction to algorithms, 2nd edition. MIT Press, Cambridge, Massachusetts.

Crozier, R. H. 1997. Preserving the information content of species: Genetic diversity, phylogeny, and conservation worth. Ann. Rev. Ecol. Syst. 28:243–268.

Crozier, R. H., L. J. Dunnett, and P.-M. Agapow. 2005. Phylogenetic biodiversity assessment based on systematic nomenclature. Evol. Bioinformatics Online 1:11–36.

Faith, D., G. Carter, G. Cassis, S. Ferrier, and L. Wilkie. 2003. Complementarity, biodiversity viability analysis, and policy-based algorithms for conservation. Environ. Sci. Policy 6:311–328.

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

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

Faith, D. P., and K. J. Williams. 2006. Phylogenetic diversity and biodiversity conservation. Pages 233–235 *in* McGraw-Hill Yearbook of Science and Technology, McGraw-Hill, New York.

Gaston, K. 1996. Species richness: Measure and measurement. Pages 77–113 *in* Biodiversity: A biology of numbers and difference (K. Gaston, ed.). Blackwell Science, Cambridge.

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

Hartmann, K., and M. Steel. 2006. Phylogenetic diversity: From combinatorics to ecology. *in* New mathematical models in evolution in preparation (O. Gascuel and M. Steel, eds.). Oxford University Press, Oxford, UK.

Johst, K., M. Drechsler, and F. Wätzold. 2002. An ecological-economic modelling procedure to design compensation payments for the efficient spatio-temporal allocation of species protection measures. Ecol. Econ. 41:37–49.

Korte, B., L. Lovaśz, and R. Shrader. 1991. Greedoids, algorithms and combinatorics. Springer, Berlin.

Lamberson, R. H., R. McKelvey, B. R. Noon, and C. Voss. 1992. A dynamic analysis of northern spotted owl viability in a fragmented forest landscape. Conserv. Biol. 6:505–512.

Lewis, L. A., and P. O. Lewis. 2005. Unearthing the molecular phylodiversity of desert soil green algae (Chlorophyta). Syst. Biol. 54:936–947.

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

Mooers, A. O., S. B. Heard, and E. Chrostowski. 2005. Evolutionary heritage as a metric for conservation. Pages 120–138 *in* Phylogeny and conservation (A. Purvis, T. Brooks, and J. Gittleman, eds.). Cambridge University Press, Cambridge, UK.
Pardi, F., and N. Goldman. 2005. Species choice for comparative genomics: No need for cooperation. PLoS Genet. 1:e71.
Redding, D. W., and A. Ø. Mooers. 2006. Incorporating evolutionary measures into conservation prioritization. Conserv. Biol. In Revision.
Reist-Marti, S., A. Abdulai, and H. Simianer. 2006. Optimum allocation of conservation funds and choice of conservation programs for a set of African cattle breeds. Genet. Select. Evol. 38:99–126.
Simianer, H., S. Marti, J. Gibson, O. Hanotte, and J. Rege. 2003. An approach to the optimal allocation of conservation funds to minimize loss of genetic diversity between livestock breeds. Ecol. Econ. 45:377–392.
Soutullo, A., S. Dodsworth, S. B. Heard, and A. O. Mooers. 2005. Distribution and correlates of carnivore phylogenetic diversity across the Americas. Anim. Conserv. 8:249–258.
Steel, M. 2005. Phylogenetic diversity and the greedy algorithm. Syst. Biol. 54:527–529.
van der Heide, C., J. C. van den Bergh, and E. C. van Ierland. 2005. Extending Weitzman's economic ranking of biodiversity protection: Combining ecological and genetic considerations. Ecol. Econ. 55:218–223.
Vane-Wright, R. I., C. J. Humphries, and P. H. Williams. 1991. What to protect? Systematics and the agony of choice. Biol. Conserv. 55:235–254.
Weitzman, M. L. 1992. On diversity. Q. J. Econ. 107:363–405.
Weitzman, M. L. 1995. Diversity functions. *in* Biodiversity loss: Economic and ecological issues (C. Perrings, K.-G. Mäler, C. Folk, C. Holling, and B.-O. Jansson, eds.). Cambridge University Press, Cambridge, UK.
Weitzman, M. L. 1998. The Noah's Ark Problem. Econometrica 66:1279–1298.
Witting, L., and V. Loeschcke. 1995. The optimization of biodiversity conservation. Biol. Conserv. 71:205–207.
Witting, L., J. Tomiuk, and V. Loeschke. 2000. Modelling the optimal conservation of interacting species. Ecol. Model. 125:123–143.

First submitted 14 December 2005; reviews returned 19 January 2006; final acceptance 22 February 2006
Associate Editor: Dan Faith

APPENDIX 1

NONLINEAR CONSERVATION EXPENDITURE AND TAXON SURVIVAL RELATIONSHIP

We describe a technique by which Generalized Noah's Ark Problems (g-NAPs) that satisfy certain conditions are transformed to equivalent NAPs. This transformation is used to show that there is one form of $g_j(q_j)$ that transforms to the type of problem considered in Scenario 1 and can therefore be solved using a greedy algorithm.

We may assume that there is some smallest unit by which the q_j can be increased or decreased (the absolute limit is the smallest unit of currency), and we denote this by δ. Recalling that the conservation budget is B, there are $m = B/\delta$ units of budget to allocate. In the transformed problem, each taxon, j, from the original g-NAP is replaced by m derived taxa. The m derived taxa are all located in the same position in the tree as the original taxon j was, this is possible as these taxa have pendant edges of zero length and the original taxon j is a leaf node.

Each of the derived taxa represents a budget unit being allocated to the original taxon j. Consequently, there is an ordering of these taxa, the first of the m taxa derived from j represents a single budget unit being allocated to j and so on. Given a solution to the transformed NAP, the corresponding solution to the g-NAP is found by noting how many derived taxa are conserved for each original taxon, j—this indicates the number of budgetary units to allocate to j.

The cost of each derived taxon is simply the cost of a single budgetary unit (δ). Next it is necessary to place some restrictions on the parameters, a_{jl} and b_{jl} of the derived taxa. Consider a taxon, j, in the original g-NAP. When the first l taxa derived from j are conserved, the probability that at least one of the taxa derived from j remains extant is:

$$z_{jl} = 1 - \prod_{o \le l}(1 - b_{jo}) \prod_{r > l}(1 - a_{jr}).$$

For the derived NAP to be equivalent to the original g-NAP, z_{jl} should equal the probability that j remains extant if $l\delta$ is spent on conserving it: $q_j(l\delta)$. For each original taxon, j, this gives $m + 1$ equations for the $2m$ parameters b_{jl} and a_{jl}:

$$z_{jl} = q_j(l\delta). \tag{8}$$

Lemma 1. *The above transformation results in a NAP that is equivalent to the original g-NAP provided that for all j and for all l:*

$$\frac{b_{j(l+1)} - a_{j(l+1)}}{1 - a_{j(l+1)}} \le \frac{b_{jl} - a_{jl}}{1 - a_{jl}} \tag{9}$$

Proof. From the derivation of the condition on a_{jl} and b_{jl} it is apparent that conserving the first l taxa derived from the original taxon j is equivalent to spending δl on conserving taxon j. However, this assumes that the derived taxa are added in the appropriate order, the remainder of this proof shows that this is guaranteed if Equation (9) is satisfied.

Consider only those taxa derived from a single taxon, j, of which the first l taxa in the sequence have been conserved. The increase in z_{jl} that the addition of one of the remaining taxa, o, will provide is:

$$\Delta z_{jl}(o) = \frac{b_{jo} - a_{jo}}{1 - a_{jo}} \prod_{r \le l}(1 - b_{jr}) \prod_{s > l}(1 - a_{js}).$$

The taxon that provides the greatest increase in z_{jl} will be the taxon picked next by the greedy algorithm. Equation (9) guarantees that $\Delta z_{jl}(o)$ will be greatest for $o = l + 1$, hence the correct taxon may be added next. There may be other taxa with an equal value of $\Delta z_{jl}(o)$; however, it is only necessary for the correct sequence of taxon additions to be a possible greedy solution. As previously noted, all solutions produced by the greedy algorithm will be optimal; hence, it suffices for one of the solutions produced by the transformed NAP to be realistic.

Theorem 3. *Problems for which $g_j(q_j)$ has the form*

$$g_j(q_j) = 1 - k^{q_j}(1 - a_j) \quad \textit{with} \quad 0 \le k \le 1, \tag{10}$$

can be transformed to a NAP of the type described in Scenario 1. Consequently, such problems can be solved using a greedy algorithm.

Proof. To satisfy the restrictions imposed on Scenario 1, the costs of each transformed taxon must be equal and Equation (4) must be satisfied. The former restriction is trivial as each taxon costs δ to conserve, the remainder of the proof shows that a transformation satisfying the latter condition exists.

The condition imposed on the transformation $[z_{jl} = g_j(l\delta)]$ for this particular $g_j(q_j)$ is:

$$1 - \prod_{o \le l}(1 - b_{jo}) \prod_{r > l}(1 - a_{jr}) = 1 - k^{l\delta}(1 - a_j). \tag{11}$$

Applying the necessary condition for the transformed NAP to be a Scenario 1 type problem (Equation (4)) this becomes:

$$1 - \kappa^l \prod_{r}(1 - a_{jr}) = 1 - k^{l\delta}(1 - a_j). \tag{12}$$

This has a simple solution, $\kappa = k^\delta$ and $a_{jr} = 1 - (1 - a_j)^{1/m}$ for all j, r. This solution also trivially satisfies Equation (9) because all taxa derived from an original taxon are identical (and hence the transformed NAP is equivalent to the original g-NAP).