Split Diversity: Measuring and Optimizing Biodiversity Using Phylogenetic Split Networks

Olga Chernomor, Steffen Klaere, Arndt von Haeseler, and Bui Quang Minh

Abstract About 20 years ago the concepts of phylogenetic diversity and phylogenetic split networks were separately introduced in conservation biology and evolutionary biology, respectively. While it has been widely recognized that biodiversity assessment should better take into account the phylogenetic tree of life, it has also been widely acknowledged that phylogenetic networks are more appropriate for phylogenetic analysis in the presence of hybridization, horizontal gene transfer, or contradicting trees among genomic loci. Here, we aim to combine phylogenetic diversity and networks into one concept, split diversity (SD), which properly measures biodiversity for conflicting phylogenetic signals. Moreover, we reformulate well-known conservation questions under the SD framework and present computational methods to solve these, in general, computationally intractable questions. Notably, integer programming, a technique widely used to solve many real-life problems, serves as a general and efficient strategy that delivers optimal solutions to many biodiversity optimization problems. We finally discuss future directions for the new concept.

Keywords Biodiversity optimization • Phylogenetic diversity • Phylogenetic split networks • Split diversity • Integer programming

O. Chernomor • A. von Haeseler
Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories,
University of Vienna, Medical University of Vienna, Vienna, Austria

Bioinformatics and Computational Biology, Faculty of Computer Science,
University of Vienna, Vienna, Austria
e-mail: olga.chernomor@univie.ac.at; arndt.von.haeseler@univie.ac.at

S. Klaere
Department of Statistics, School of Biological Sciences,
University of Auckland, Auckland, New Zealand
e-mail: s.klaere@auckland.ac.nz

B.Q. Minh (✉)
Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories,
University of Vienna, Medical University of Vienna, Vienna, Austria
e-mail: minh.bui@univie.ac.at

R. Pellens, P. Grandcolas (eds.), *Biodiversity Conservation and Phylogenetic Systematics*, Topics in Biodiversity and Conservation 14,
DOI 10.1007/978-3-319-22461-9_9

Introduction

The previous book chapters show that in the presence of phylogenetic information it is more appropriate to assess biodiversity based on phylogenetic trees than on the concept of species richness (see also May 1990; Vane-Wright et al. 1991). *Phylogenetic diversity* (PD; Faith 1992) is a popular measure of the amount of evolutionary history encompassed by the species under consideration. Given a phylogenetic tree for a set of taxa, PD of a taxon subset is defined as the sum of the branch lengths of the minimal subtree connecting those taxa. The definition of PD per se requires "a reliable estimate of phylogenetic relationships among the taxa" (Faith 1992). However, such a reliable estimate is sometimes hard to obtain due to, for example, model misspecification (Jermiin et al. 2008) or even intrinsically non-treelike evolutionary patterns. More recently, phylogenomic studies often revealed conflicting phylogenetic signals among genomic loci, adding the complication how to compute PD from multiple trees.

Figure 1 illustrates the problem. Here, phylogenetic trees are reconstructed for ten pheasant species from the mitochondrial cytochrome *b* gene (CYB) and the intron 3 of the dimerization cofactor of hepatocyte nuclear factor 1 (DCoH3) (data from Kimball and Braun 2008). The two resulting trees, denoted by T_{CYB} and T_{DCoH3}, clearly separate the two genera Gallus (junglefowl) and Polyplectron (peacock-pheasant). However, they strongly contradict within the Gallus clade. For example, *G. sonneratii* (grey junglefowl) and *G. varius* (green junglefowl) are the basal Gallus species in T_{CYB} and T_{DCoH3}, respectively. The trees also disagree on the phylogenetic positions of *P. emphanum* (Palawan peacock-phesant) and *P. malacense* (Malayan peacock-pheasant). Moreover, edge lengths of the trees represented by

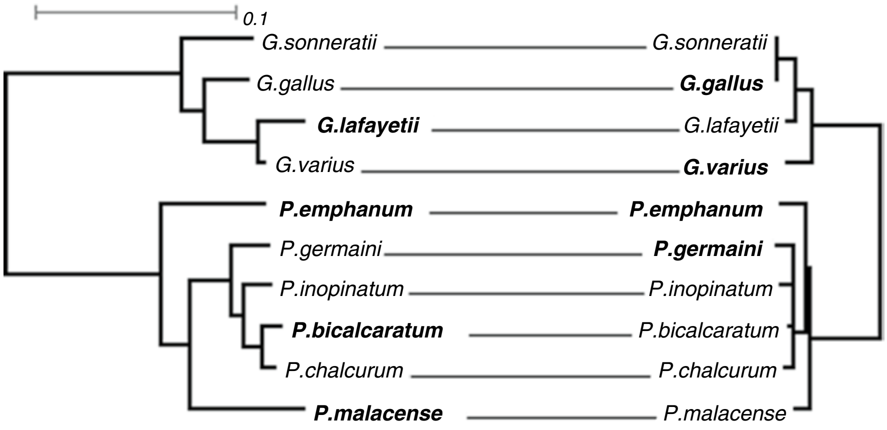

Fig. 1 Maximum likelihood phylogenetic trees inferred with IQ-TREE (Minh et al. 2013) from the mitochondrial CYB and the nuclear intron DCoH3 for four Gallus (junglefowl) and six Polyplectron (peacock-pheasant) species. The *scalebar* represents the expected number of nucleotide substitutions per site. Highlighted in *boldface* are the four species maximizing phylogenetic diversity

the expected numbers of substitutions per site substantially differ between the trees. This particular example reflects the fact that the evolutionary relationships among these birds are still controversial and more data is needed to elucidate the galliform tree of life (e.g., Wang et al. 2013).

If one is interested in selecting four species maximizing PD, then one indeed ends up with two different sets of species (highlighted in bold-face, Fig. 1) and only *P. emphanum* occurs in both subsets.

To resolve this issue, we introduced the concept of *Split Diversity* (SD), which generalizes PD by combining information from multiple trees (Minh et al. 2009). For example, SD of a taxon set can be defined as the average PD of the two trees. By maximizing SD one then simultaneously maximizes PDs over all trees, which captures conflicting phylogenetic signals between the trees. Moreover, computing SD this way is equivalent to computing "phylogenetic diversity" from the so-called *phylogenetic split networks* (Bandelt and Dress 1992a; Huson et al. 2010). SD has also been recently applied to prioritize populations for conservation (Volkmann et al. 2014). In the following we formalize the concept of split networks and the measure of split diversity. Further, we reformulate well-known biodiversity optimization problems under the framework of SD, present algorithmic solutions and computational tools to these problems. Finally conclude the chapter with future perspectives.

Phylogenetic Split Networks

Rooted phylogenetic trees as shown in Fig. 1 are well understood. Here, both trees show that the common ancestor of the taxa considered has the ancestors of the two genera as direct descendants. In general, interior nodes indicate ancestral taxa of the leaf nodes, and the edge lengths give an estimate of the amount of change observed between nodes. However, if one wishes to combine the information in both trees, it becomes difficult to identify clear ancestors. For example, T_{CYB} and T_{DCoH3} disagree whether *G. sonneratii* or *G. varius* is the basal Gallus species. In order to visualize these conflicts phylogenetic split networks have been devised.

We start by describing splits. A *split*, denoted by $A|B$, is defined as a bipartition of the taxon set X into two disjoint subsets A and B, indicating that there is an observable amount of divergence between the two subsets. Every edge in a tree generates a split. If one removes an edge, the tree decomposes into two subtrees, each of which connects a unique set of leaves. T_{CYB} has 17 splits (edges), while T_{DCoH3} has 15 splits (2 splits in T_{DCoH3} have zero length and are collapsed as they do not influence subsequent computations). Figure 2a shows the union set Σ of 20 distinct splits occurring in the pheasant trees (Fig. 1). T_{CYB} and T_{DCoH3} share the ten *trivial splits* $\sigma_1, \sigma_2, \ldots, \sigma_{10}$ corresponding to external edges of the trees. The trees also share two non-trivial splits σ_{13} and σ_{16}, where σ_{16} corresponds to the internal edges separating Gallus from Polyplectron species. The remaining splits are unique to each tree.

a

	Species → Splits ↓	*P. bicalcaratum*	*P. chalcurum*	*P.emphanum*	*P. germaini*	*P.inopinatum*	*P. malacense*	*G. gallus*	*G. lafayetii*	*G. sonneratii*	*G. varius*
Trivial splits	σ_1	*	.	.	.	.	.	.	.	.	.
	σ_2	.	*	.	.	.	.	.	.	.	.
	σ_3	.	.	*	.	.	.	.	.	.	.
	σ_4	.	.	.	*	.	.	.	.	.	.
	σ_5	.	.	.	.	*	.	.	.	.	.
	σ_6	.	.	.	.	.	*	.	.	.	.
	σ_7	.	.	.	.	.	.	*	.	.	.
	σ_8	.	.	.	.	.	.	.	*	.	.
	σ_9	.	.	.	.	.	.	.	.	*	.
	σ_{10}	.	.	.	.	.	.	.	.	.	*
Non-trivial splits	σ_{11}	*	*	.	.	.	.	.	.	.	.
	σ_{12}	*	*	.	.	*	.	.	.	.	.
	σ_{13}	*	*	.	*	*	.	.	.	.	.
	σ_{14}	*	*	.	*	*	*	.	.	.	.
	σ_{15}	*	*	*	*	*	.	.	.	.	.
	σ_{16}	.	.	.	.	.	.	*	*	*	*
	σ_{17}	.	.	.	.	.	.	*	*	.	*
	σ_{18}	.	.	.	.	.	.	.	*	.	*
	σ_{19}	.	.	.	.	.	.	*	*	*	.
	σ_{20}	.	.	.	.	.	.	*	.	*	.

b

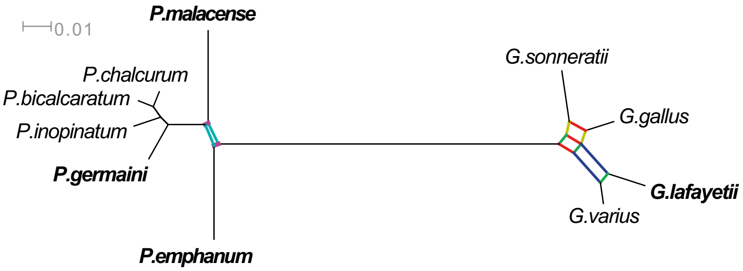

Fig. 2 (**a**) Set of all splits extracted from the trees in Fig. 1. Each split σ is a bipartition $A|B$, where '*' and '.' represent taxa in A and B, respectively. Conflicting splits are colored. (**b**) Visualization of this split set as a phylogenetic split network. Conflicting splits are colored accordingly and depicted by parallelograms. Here, split weights are assigned as the mean of the weight of the corresponding edges in the two trees. Highlighted in *boldface* are the four species maximizing split diversity

This split set is visualized in a phylogenetic split network (Fig. 2b). The major difference to trees is that the interior nodes of a split network cannot be regarded as representing ancestral taxa. Instead, the weight of a split $A|B$ indicates the amount of difference between the taxon set A and B. A split is visualized by a single edge or a set of parallel edges. The former indicates that the split does not conflict any other splits, while the latter indicates at least one conflict. Therefore, two conflicting splits are visualized by a parallelogram. For example, σ_{14} (in cyan, Fig. 2) and σ_{15} (in pink) contradict each other on the placement of *P. emphanum* and *P. malacense*. This disagreement generates a narrow parallelogram at the basal Polyplectron.

If more than two splits are in disagreement, the split network will show multiple connected parallelograms. For example, σ_{17} (in red, Fig. 2) conflicts with σ_{19} (in green) and σ_{20} (in yellow). σ_{19} also contradicts σ_{18} (in blue). Therefore, $\sigma_{17}, \sigma_{18}, \sigma_{19}$ and σ_{20} are visualized by three red, two blue, three green, and two yellow parallel edges, respectively. This generates three parallelograms within Gallus (Fig. 2b).

Not every split set can be visualized in two dimensions. For example, assuming that we had a third tree that places *G. gallus* at the basal Gallus lineage. This would introduce one split contradicting with both σ_{17} and σ_{19}. These triple-wise conflicting splits are depicted by a three dimensional parallelepiped. The resulting split network is not easily visualized anymore. However, for the following it suffices to directly work on the split set (Fig. 2a).

The Measure of Split Diversity

Given a split set Σ, the SD of a taxon subset Y is defined as the sum of the weights λ of all splits separating taxa in Y. Here, a split $A \mid B \in \Sigma$ *separates* Y if $Y \cap A$ and $Y \cap B$ are both non-empty. Thus, we get

$$SD(Y) = \sum_{\sigma \in \Sigma : \sigma \text{ separates } Y} \lambda_\sigma$$

To illustrate, given Σ in Fig. 2, for Y={*P. malacense*, *P. germaini*, *P. emphanum*, *G. lafayetii*} we have $SD(Y) = \lambda_3 + \lambda_4 + \lambda_6 + \lambda_8 + \lambda_{13} + \lambda_{14} + \ldots + \lambda_{19}$, where $\lambda_i = \lambda_{\sigma_i}$ is defined as the average of the corresponding branch lengths in T_{CYB} and T_{DCoH3}. Here, contradicting splits such as σ_{17} and σ_{19} are considered in the SD computation.

If the split set Σ corresponds to a tree (i.e. no conflicting splits exist in Σ), then SD is equivalent to PD. The definition of SD therefore generalizes PD. For this reason we focus on SD for the remaining of the chapter.

Biodiversity Optimization Problems

Conservation problems mainly fall into two categories: taxon selection and reserve selection (Fig. 3), where the conservation targets are either taxa or geographical areas, respectively. Under PD, the simplest taxon selection problem (Faith 1992) is

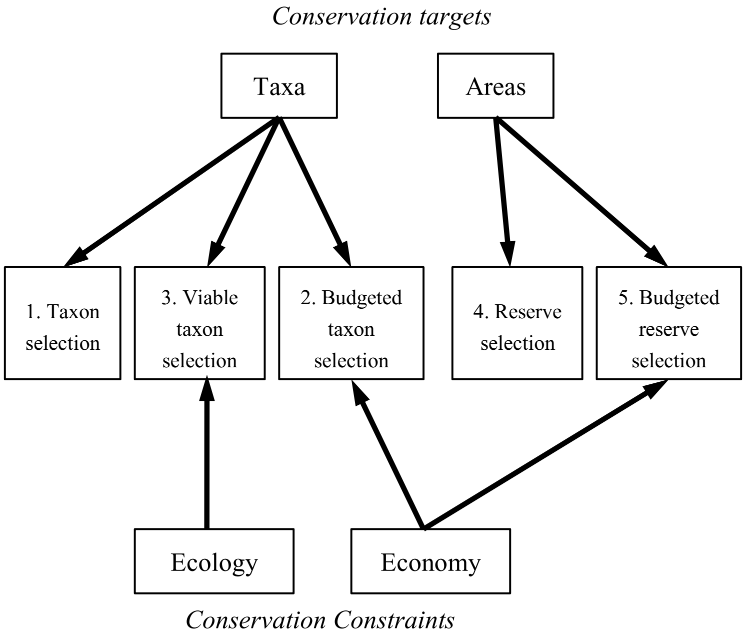

Fig. 3 The "network" of biodiversity optimization problems

to identify a subset of k taxa that maximizes PD on a phylogenetic tree of n taxa ($2 \leq k < n$). For reserve selection we define PD on a subset of areas as the PD of the union taxon set of the areas. The simplest reserve selection problem is analogously to identify a subset of k areas that maximizes PD over all subsets of k areas. In the following, we reformulate these problems using SD and further integrate economical and ecological constraints into the extensions.

Taxon Selection Problems

We start with the simplest taxon selection problem formally defined as:

Problem 1 (Taxon Selection)
Given a phylogenetic split set for n taxa, find a subset of k taxa that maximizes SD over all subsets of k taxa.

As an illustration, given the split set for ten pheasants (Fig. 2) we want to select four taxa maximizing SD. By doing so we yield an optimal subset (highlighted in bold-face; Fig. 2b), which shares three taxa (*P. emphanum*, *P. malacense*, and *G. lafayetii*) with the CYB-based subset (left panel of Fig. 1) and only two taxa (*P. emphanum* and *P. germaini*) with DCoH3-based subset (right panel of Fig. 1). The SD approach therefore provides a "consensus" solution over the two independent PD analyses. Problem 1 is known to be NP-hard (Spillner et al. 2008), which means that to find an optimal set it may, in the worst case, necessary to compute the SD for the exponentially many subsets *n*.

Problem 1 implicitly assumes that each taxon requires the same amount of resources for conservation. If we knew the preservation costs for each taxon and were provided with a finite budget, then a more realistic scenario is to allocate this budget among the taxa so as to obtain the highest diversity. This process is known as conservation triage (Bottrill et al. 2008) and formally defined as:

Problem 2 (Budgeted Taxon Selection)
Given a split set and conservation costs for each taxon, find a subset of taxa whose total conservation costs do not exceed a predefined budget while maximizing SD.

Problem 1 and 2 ignore ecological relationships between taxa. In real life species interact with each other within a dependency network such as predator-prey relationships (Witting et al. 2000; van der Heide et al. 2005; Moulton et al. 2007). In general, a dependency network is, typically, an acyclic directed graph, where nodes in the graph represent taxa and edges represent dependencies between nodes. Figure 4 shows an artificial example of such a network for the pheasants. Here, *G. sonneratii* depends on *P.malacense* and *P.germaini*, depicted by two edges connecting *G.sonneratii* with these two taxa. We note that this is a purely fictional example, but it illustrates the major principles of including a dependency structure in conservation decisions.

A taxon is called *viable in a subset* of taxa if this taxon does not depend on any other taxon, or if it does depend on some taxa, then at least one of them is also present in the subset. For example, *G.sonneratii* is viable in a subset if this subset also contains *P.malacense* or *P.germaini*. *P.emphanum* and *G.gallus* are viable in any (sub)set since they do not depend on any other species.

A subset is called viable if all its taxa are viable in this set. For example, {*P. emphanum*, *P. bicalcaratum*, *P. germaini*, *G. sonneratii*} is a viable subset, whereas {*P.emphanum*, *P.bicalcaratum*, *G.lafayetii*, *G.sonneratii*} is not viable.

We now formally define the viable taxon selection problem as

Problem 3 (Viable Taxon Selection)
Given a split set and a dependency network, find a viable subset of *k* taxa, which maximizes SD.

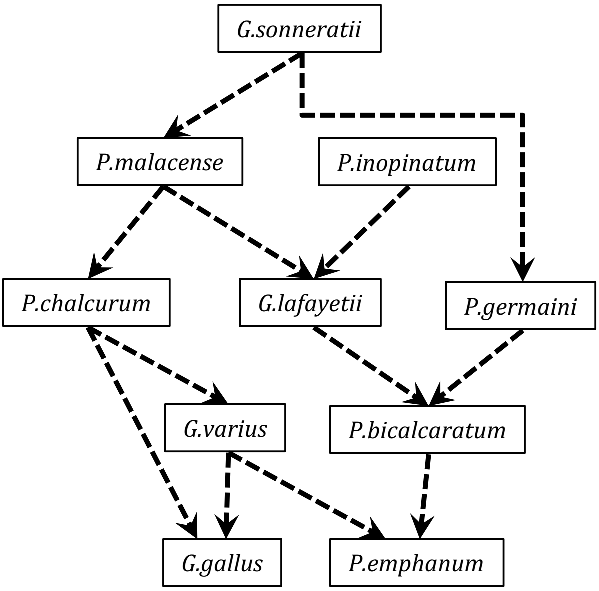

Fig. 4 Artificial example of dependency network for the pheasant data set

Reserve Selection Problems

For reserve selection we define the SD of a subset of areas as the SD of the union set of taxa present in these areas. The reserve selection is formalized as:

> ***Problem 4 (Reserve Selection)***
> Given a split set for n taxa distributed in m areas, find a subset of k areas that maximizes SD over all subsets of k areas.

To illustrate the problem consider the geographical distribution of the ten pheasants (Table 1). The data were obtained from the global biodiversity information facility (www.gbif.org; accessed on December 1st, 2013), where a country is listed as habitat only if there are at least three observations for the species. Table 1 shows that these pheasants occur in eight countries in South Asia. *G. gallus* and *P. bicalcaratum* occur in seven and two countries, respectively, whereas the remaining species are endemic to one country. Indonesia and Malaysia each host three species, Sri Lanka only one species, and the remaining five countries are home to two species each.

If one wants to select four countries with maximal diversity, then the decision heavily depends on the trees or network (Figs. 1 and 2b). Table 2 shows that using

Table 1 Presence/absence of ten pheasants in eight countries obtained from the global biodiversity information facility (http://www.gbif.org)

Species	Bhutan, BT	Indonesia, ID	India, IN	Sri Lanka, LK	Philippines, PH	Malaysia, MY	Thailand, TH	Vietnam, VN
P. bicalcaratum	x						x	
P. chalcurum		x						
P. emphanum					x			
P. germaini								x
P. inopinatum						x		
P. malacense						x		
G. gallus	x	x	x		x	x	x	x
G. lafayetii				x				
G. sonneratii			x					
G. varius		x						

Table 2 Four countries maximizing PD on the CYB tree (first column), PD on the DCoH3 tree (second column), and SD on the split network (third column)

PD – CYB	PD – DCoH3	SD
Malaysia (3)	Indonesia (3)	**Malaysia** (3)
Philippines (2)	**Malaysia** (3)	**Philippines** (2)
Sri Lanka (1)	**Philippines** (2)	Indonesia (3)
India (2)	Vietnam (2)	India (2)

Highlighted in boldface are the countries present in all optimal sets. The number of species present in the country is given in brackets

the CYB and DCoH3 regions, the optimal sets only overlap in two countries: Malaysia and Philippines. If we now maximize SD instead, then the optimal set includes these two countries, the third one preferred by the PD-DCoH3 set (Indonesia), and the fourth one by the PD-CYB set (India). The union of the species sets for the selected areas contains seven species.

If budget data is available, then we have a budgeted reserve selection problem. Here, preserving these species in each country comes at a cost and we need to select those countries that maximize SD within an allocated budget.

Problem 5 (Budgeted Reserve Selection)
Given a split set for n taxa distributed on m areas and conservation costs for each area, find a subset of areas whose total conservation costs do not exceed a predefined budget while maximizing SD.

Computational Methods in Conservation Planning

The algorithms to solve the aforementioned Problems 1–5 are those that are guaranteed to produce an optimal solution, often referred to as exact algorithms, and those that are not. The former includes algorithms that are based on integer programming and dynamic programming, whereas the latter comprise greedy algorithms, approximation algorithms and algorithms based on simulated annealing. We will start with greedy algorithms, as they are simple and probably most widely applied in conservation planning.

Greedy Algorithms

Greedy algorithms are a simple and general heuristic strategy but, usually, do not guarantee optimal solutions. Kirkpatrick (1983) was probably the first to apply a greedy algorithm to find a solution to Problem 4, the simple reserve selection, but under the species richness concept. His greedy algorithm coined “complementarity principle” first identifies the most species-rich area. In the second step, it finds the area, which “adds” the most numbers of new species to the firstly chosen area. This is repeated until *k* areas are obtained. Such a complementarity principle has been applied to maximize PD (Faith 1992) and also applied elsewhere (e.g., Vane-Wright et al. 1991; Pressey et al. 1997). Recently, Bordewich and Semple (2008) have proven that the greedy algorithm applied to Problem 5 under PD will generate a solution that has at least ~63 % of the PD of the optimal solution, which is the best possible approximation ratio.

The only case, where a greedy algorithm delivers the optimal solution is the taxon selection (Problem 1) under PD on trees (Pardi and Goldman 2005; Steel 2005). An efficient implementation of such a greedy algorithm (Minh et al. 2006) finds a solution for trees with millions of taxa within seconds on a standard PC. Greedy algorithms have been further examined in conservation biology (Moulton et al. 2007; Bordewich et al. 2008).

Obviously greedy algorithms can be applied for Problems 1–5 to maximize SD. The general idea is to start with one target (either taxon or area) having the highest SD. We then choose the second target “adding” the most SD while still satisfying the constraints (budget or viability constraints). We repeat this step until no further target can be added (e.g., exceeding *k* targets for Problem 1, 3, and 4 or exceeding the budget for Problem 2 and 5). As an illustration the greedy algorithm is applied for Problem 4 to find four countries showing the highest pheasant SD for the split network (Fig. 2b) and known geographical distribution (Table 1) as follows. Malaysia is first selected as it contains the highest SD. Philippines, Indonesia, and India are selected in the next steps. In this particular example the greedy algorithm happens to obtain the optimal set of four countries (Table 2).

Integer Programming

Integer Programming (IP; Dantzig et al. 1954; Gomory 1958) is a widely used and powerful optimization technique to solve a variety of decision-making problems (Wolsey 1998; Jünger et al. 2010). IP methods maximize or minimize a linear objective function subject to linear constraints (equalities or inequalities) when one or more variables are restricted to be integers. Theoretically solving IP is NP-hard. However, thanks to powerful solvers like CPLEX (2012) and GUROBI (Gurobi Optimization Inc. 2013), problems with thousands of variables and constraints can be solved optimally within reasonable time (Jünger et al. 2010; and references therein).

The first application of IP in conservation problems goes back to (Cocks and Baird 1989), who solved the reserve selection (Problem 4) under species richness. Such IP formulations have been extended to more realistic scenarios (Underhill 1994; Church et al. 1996; Possingham et al. 2000), to maximize PD (Rodrigues and Gaston 2002; Rodrigues et al. 2005), and to maximize SD (Minh et al. 2010).

Here, we show how to model biodiversity optimization problems 1–5 in IP parlance, which allows available IP software packages to solve the problem. We first introduce some notations and definitions and further exemplify IP formulations for Problems 1–5 using the pheasant data set.

IP for Taxon Selection Problems

Given a set of n taxa, we encode a subset S by an n-element binary vector with entries of 0 and 1 indicating the absence and presence of the corresponding taxa in S. The elements of this vector are called taxon variables. For the pheasant data set there are ten taxon variables x_{PB}, x_{PC}, x_{PE}, x_{PG}, x_{PI}, x_{PM}, x_{GG}, x_{GL}, x_{GS}, x_{GV} (indices follow initials of species names). We say that a split $\sigma = A|B$ is preserved in S if A and B each contain at least one taxon from S. For each split σ we introduce a binary split variable y_σ, where $y_\sigma = 1$ if σ is *preserved* in S and 0 otherwise.

Each y_σ is fully identified from taxon variables by two split constraints as follows. σ_1 is a trivial split that separates *P. bicalcaratum* from the remaining taxa. σ_1 is preserved (i.e. $y_1 = 1$) if *P. bicalcaratum* and at least another taxon are preserved (see Fig. 2a for the definition of the splits). This condition is expressed by two inequalities:

$$y_1 \le x_{PB}, y_1 \le x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GG} + x_{GL} + x_{GS} + x_{GV}$$

In fact, the second inequality always holds because $k \ge 2$ and thus is ignored. Now consider the non-trivial split σ_{17}, which separates *G. gallus*, *G. lafayetii*, and *G. varius* from the remaining taxa. σ_{17} is preserved if at least one of *G. gallus*, *G. lafayetii*, and *G. varius* and one of the remaining taxa are preserved. Therefore,

$$y_{17} \le x_{GG} + x_{GL} + x_{GV}, y_{17} \le x_{PB} + x_{PC} + x_{PI} + x_{PG} + x_{PE} + x_{PM} + x_{GS}$$

The remaining split constraints are listed in Table 3.

Based on split variables one can rewrite SD of S as:

$$SD(S) = \sum_{\sigma} \lambda_\sigma y_\sigma \tag{1}$$

where λ_σ is the weight of split σ. This is the objective function that we want to maximize for all problems (1–5).

In the taxon selection Problem 1 the size of an optimal subset is constrained by a predefined number k, meaning that:

$$x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GG} + x_{GL} + x_{GS} + x_{GV} \le k \tag{2}$$

We also require that taxon and split variables are binary

$$x_i \in \{0,1\},\ \forall \text{ taxon } i \tag{3}$$

$$y_\sigma \in \{0,1\},\ \forall \text{ split } \sigma \tag{4}$$

IP Formulation of Problem 1
Maximize objective function (1), subject to subset size constraint (2), binary constraints (3, 4), split constraints (5) (see Table 3).

Suppose we are given a total budget B. Let c_i denote conservation costs for taxon i. We can then substitute constraint (2) by the budget constraint

$$\sum_i c_i x_i \le B \tag{6}$$

Together with previous constraints we have the IP formulation of Problem 2 by:

IP Formulation for Problem 2
Maximize objective function (1), subject to budget constraint (6), binary constraints (3, 4), and split constraints (5) (Table 3).

We now model viability constraints that operate on taxon variables as follows. *G. sonneratii* depends on *P.malacense* and *P.germaini* (Fig. 4). Therefore, the viability constraint for *G. sonneratii* is simply

$$x_{PM} + x_{PG} \ge x_{GS}$$

This ensures that x_{GS} is 1 (i.e., *G. sonneratii* is selected for conservation) only if at least one of x_{PM} and x_{PG} is also 1. Viability constraints for all the other taxa are listed in Table 3. Now, the IP formulation for viable taxon selection can be obtained by simply including viability constraints to Problem 1:

IP Formulation of Problem 3
Maximize objective function (1), subject to subset size constraint (2), binary constraints (3, 4), split constraints (5), and viability constraints (7) (Table 3).

IP for Reserve Selection Problems

For reserve selection we encode a subset W of m areas by a binary vector $(z_1,z_2,\ldots,z_m)$, where z_r is 1 if area r is present in W, and 0 otherwise. We call z_r area variables. For the pheasant habitat (Table 1) we have eight area variables z_{ID}, z_{LK}, z_{BT}, z_{IN}, z_{PH}, z_{MY},

z_{TH}, z_{VN} (indices follow two-letter country codes). We now redefine split constraints in terms of area variables instead of taxon variables as follows.

Split σ_{18}, which separates *G. lafayetii* and *G. varius* from the others, is preserved if (1) *G. lafayetii* or *G. varius* is preserved and (2) at least one of the remaining taxa is preserved. Because *G. lafayetii* or *G. varius* occur in Indonesia and Sri Lanka, condition 1 is equivalent to:

$$y_{18} \le z_{ID} + z_{LK}$$

Similarly condition 2 is equivalent to:

$$y_{18} \le z_{BT} + z_{ID} + z_{IN} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$$

since the remaining taxa are found in all areas except Sri Lanka. Such area-split constraints for all other splits are listed in Table 4.

The subset size constraint has to be rewritten for countries:

$$z_{ID} + z_{LK} + z_{BT} + z_{IN} + z_{PH} + z_{MY} + z_{TH} + z_{VN} \le k \tag{8}$$

We keep binary constraints for split variables and also include such for area variables

$$z_r \in \{0,1\} \forall \text{ area } r \tag{9}$$

Reserve selection problem is then formulated as follows:

IP Formulation of Problem 4
Maximize objective function (1), subject to subset size constraint (8), binary constraints (4, 9), and area-split constraints (10) (Table 4).

For budgeted reserve selection we are given a total budget B. Let c_{ID}, c_{LK}, c_{BT}, c_{IN}, c_{PH}, c_{MY}, c_{TH}, c_{VN} denote conservation costs for each country. Then a budget constraint for areas is

$$\sum_r c_r z_r \le B \tag{11}$$

To obtain the IP formulation for Problem 5 we simply substitute subset size constraint (8) by the budget constraint (11).

IP Formulation of Problem 5
Maximize objective function (1), subject to budgetary constraint (11), binary constraints (4, 9), and area-split constraints (10) (Table 4).

Other Algorithms

While greedy algorithms and IP are general strategies for all Problems 1–5, other algorithms have been applied to solve special cases. For example, simulated annealing algorithms (Possingham et al. 2000) were introduced to solve the reserve selection Problems 4 and 5 under species richness with an opportunity to minimize the connectivity between the areas such as the boundary lengths. Dynamic programming algorithms (DPA) have been applied to solve Problem 2 under PD (Pardi and Goldman 2007). DPA was further extended to maximize SD on circular split networks (Minh et al. 2009a, b). Other special types of split networks were exploited to solve Problem 1 (Spillner et al. 2008; Bordewich et al. 2009).

Computer Software

Conservation planning software like *Marxan* (Ball et al. 2009) and *Zonation* (Moilanen et al. 2009) mainly focus on species richness. However, both programs can indirectly account for phylogenetic diversity (see also Silvano, Valdujo and Colli, chapter "Priorities for Conservation of the Evolutionary History of Amphibians in the Cerrado" and Arponen and Zupan, chapter "Representing Hotspots of Evolutionary History in Systematic Conservation Planning for European Mammals"). Only a few programs explicitly allow to compute phylogenetic diversity (Webb et al. 2008; Kembel et al. 2010). In the following we describe two programs relevant for the SD analysis.

SplitsTree

SplitsTree (Huson and Bryant 2006) is a user-friendly and leading software to reconstruct and visualize phylogenetic networks from multiple sequence alignments, distance matrices, or sets of trees. SplitsTree implements a wide range of split network inference methods such as split decomposition (Bandelt and Dress 1992b) and neighbor-net (Bryant and Moulton 2004). SplisTree has a limited ability to compute PD and SD. It works for all major platforms including Windows, Mac OS X, and Unix. More information about SplitsTree is available at http://www.splitstree.org.

PDA: Phylogenetic Diversity Analyzer

PDA (Minh et al. 2009) is a software tool that computes and maximizes species richness, PD, and SD given a variety of user-defined constraints including budget, ecological, and geographical constraints. PDA can be used in conjunction with SplitsTree to work with SD. It solves all Problems 1–5 by greedy algorithms, dynamic programming, and integer programming methods. Moreover, it supports weighted dependency networks for viable taxon selection and spatial reserve selection problems (Chernomor et al. 2015). Among other features is the computation of PD/SD endemism and complementarity (Faith et al. 2004). PDA is available as a command-line program for Windows, Mac OS X, and Unix as well as an online web service. More information about PDA is available at http://www.cibiv.at/software/pda.

Conclusions and Perspectives

In this chapter we have presented the concept of split diversity, a generalization of PD to account for contradicting phylogenetic information in biodiversity optimization. We demonstrated the new concept with a small pheasant data set. We note that this example is not realistic because neither genera are vulnerable nor the selection of entire countries is reasonable. Moreover, genetic data for galliforms are available for more genera and genomic loci (Wang et al. 2013) and the methodology developed here is well applicable to this new data.

We then presented computational tools to perform the analysis under the SD framework. Both greedy algorithms and IP can be generally applied to solve the same conservation questions, where the former quickly computes a solution and the latter ensures optimal solutions. Moreover, IP works well for data set sizes usually encountered in real data. For example, we have recently applied IP to solve the viable taxon selection (Problem 3) for 242 marine species of Caribbean coral community and the budgeted reserve selection (Problem 5) for the Cape of South Africa with 735 plant genera (Chernomor et al. 2015). IP always returned optimal sets of taxa and areas within seconds to a few minutes.

SD can be extended to include species extinction risks as developed for PD (Weitzman 1992; Witting and Loeschcke 1995). Such a "probabilistic" PD approach (see chapters "The Value of Phylogenetic Diversity" and "Reconsidering the Loss of Evolutionary History: How Does Non-random Extinction Prune the Tree-of-Life?") predicts future diversity given the fact that some species might become extinct in, say, 20 years. The problem, previously coined the Noah's Ark Problem (NAP; Weitzman 1998), is then to maximize future PD given limited budgets. The same concept can be applied to SD as follows. One first computes "survival probabilities" for each split in split networks in the same fashion as for branches in phylogenetic trees. The future SD is then defined as the dot product of the split weights and split survival probabilities. This definition of future SD consistently generalizes that of future PD.

From a computational view point, solving the extended NAP under future SD is NP-hard as proven for PD (Hartmann and Steel 2006). Dynamic programming algorithms (DPA) optimally solve the NAP under future PD in a special scenario, where the species extinction probability becomes 0 if it is given enough resources (Pardi and Goldman 2007). For general scenarios Hickey et al. (2008) devised such a DPA that gives an approximation ratio of nearly 1 compared to the optimal solution. More recently, Billionnet (2013) presented an IP approach for the NAP that runs within a few minutes for simulated 4,000-taxon cases and provides near-optimal solutions, which are only 1.2 % away from the optimal solution. It will be interesting to investigate how such DPA and IP approaches can be adapted to solve the NAP under the more general SD framework.

Acknowledgments The authors thank Tung Lam Nguyen for developing PDA web service. The authors also thank Roseli Pellens and Philippe Grandcolas for inviting us to write this chapter and three anonymous reviewers for constructive comments. This work was supported by University of Vienna (Initiativkolleg I059-N) to O.C. and the Austrian Science Fund – FWF (I760-B17) to B.Q.M. and A.v.H.

Appendix

Table 3 Objective function and constraints of taxon selection problems for the pheasant example

Maximize	$\lambda_1 y_1 + \ldots + \lambda_{20} y_{20}$	(1)
Subject to		
Size constraint:	$x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GG} + x_{GL} + x_{GS} + x_{GV} \leq k$	(2)
Binary constraints:	$x_i \in \{0,1\} \forall$ taxon i	(3)
	$y_\sigma \in \{0,1\}\ \forall \sigma = 1,..,20$	(4)
Split constraints:	$y_i \leq x_i \forall$ taxon i	(5)
	$y_{11} \leq x_{PB} + x_{PC}$	
	$y_{11} \leq x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GG} + x_{GL} + x_{GS} + x_{GV}$	
	$y_{12} \leq x_{PB} + x_{PC} + x_{PI}$	
	$y_{12} \leq x_{PE} + x_{PG} + x_{PM} + x_{GG} + x_{GL} + x_{GS} + x_{GV}$	
	$y_{13} \leq x_{PB} + x_{PC} + x_{PG} + x_{PI}$	
	$y_{13} \leq x_{PE} + x_{PM} + x_{GG} + x_{GL} + x_{GS} + x_{GV}$	
	$y_{14} \leq x_{PB} + x_{PC} + x_{PG} + x_{PI} + x_{PM}$	
	$y_{14} \leq x_{PE} + x_{GG} + x_{GL} + x_{GS} + x_{GV}$	
	$y_{15} \leq x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI}$	
	$y_{15} \leq x_{PM} + x_{GG} + x_{GL} + x_{GS} + x_{GV}$	
	$y_{16} \leq x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM}$	
	$y_{16} \leq x_{GG} + x_{GL} + x_{GS} + x_{GV}$	
	$y_{17} \leq x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GS}$	
	$y_{17} \leq x_{GG} + x_{GL} + x_{GV}$	
	$y_{18} \leq x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GG} + x_{GS}$	
	$y_{18} \leq x_{GL} + x_{GV}$	
	$y_{19} \leq x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GV}$	
	$y_{19} \leq x_{GG} + x_{GL} + x_{GS}$	
	$y_{20} \leq x_{PB} + x_{PC} + x_{PE} + x_{PG} + x_{PI} + x_{PM} + x_{GL} + x_{GV}$	
	$y_{20} \leq x_{GG} + x_{GS}$	

(continued)

Table 3 (continued)

Budget constraint:	$c_{PB}x_{PB} + c_{PC}x_{PC} + \ldots + c_{GV}x_{GV} \le B$	(6)
Viability constraints:	$x_{PB} \le x_{PE}$	(7)
	$x_{PC} \le x_{GG} + x_{GV}$	
	$x_{PG} \le x_{PB}$	
	$x_{PI} \le x_{GL}$	
	$x_{PM} \le x_{PC} + x_{GL}$	
	$x_{GL} \le x_{PB}$	
	$x_{GS} \le x_{PM} + x_{PG}$	
	$x_{GV} \le x_{GG} + x_{PE}$	

Table 4 Objective function and constraints of reserve selection problems for the pheasant example. Due to the fact that *G. gallus* is contained in all but one area there are many area-split constraints of the form $y_\sigma \le z_{BT} + z_{ID} + z_{IN} + z_{LK} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$. Such constraints are redundant since $k \ge 2$, and thus omitted

Maximize	$\lambda_1 y_1 + \ldots + \lambda_{20} y_{20}$	(1)
Subject to		
Size constraint:	$z_{ID} + z_{LK} + z_{BT} + z_{IN} + z_{PH} + z_{MY} + z_{TH} + z_{VN} \le k$	(8)
Binary constraints:	$z_r \in \{0,1\} \forall$ area r	(9)
	$y_\sigma \in \{0,1\}\ \forall \sigma = 1,..,20$	(4)
Area-split constraints:	$y_1 \le z_{BT} + z_{TH}$	(10)
	$y_2 \le z_{ID}$	
	$y_3 \le z_{PH}$	
	$y_4 \le z_{VN}$	
	$y_5 \le z_{MY}$	
	$y_6 \le z_{MY}$	
	$y_7 \le z_{BT} + z_{ID} + z_{IN} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
	$y_8 \le z_{LK}$	
	$y_9 \le z_{IN}$	
	$y_{10} \le z_{ID}$	
	$y_{11} \le z_{BT} + z_{ID} + z_{TH}$	
	$y_{12} \le z_{BT} + z_{ID} + z_{MY} + z_{TH}$	
	$y_{13} \le z_{BT} + z_{ID} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{14} \le z_{BT} + z_{ID} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{15} \le z_{BT} + z_{ID} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{16} \le z_{BT} + z_{ID} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{17} \le z_{BT} + z_{ID} + z_{IN} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{18} \le z_{BT} + z_{ID} + z_{IN} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{18} \le z_{ID} + z_{LK}$	
	$y_{19} \le z_{BT} + z_{ID} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{20} \le z_{BT} + z_{ID} + z_{LK} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
	$y_{20} \le z_{BT} + z_{ID} + z_{IN} + z_{PH} + z_{MY} + z_{TH} + z_{VN}$	
Budget constraint:	$c_{ID} z_{ID} + c_{LK} z_{LK} + \ldots + c_{VN} z_{VN} \le B$	(11)

References

Ball IR, Possingham HP, Watts M (2009) Marxan and relatives: software for spatial conservation prioritisation. In: Moilanen A, Wilson KA, Possingham HP (eds) Spatial conservation prioritisation: quantitative methods and computational tools. Oxford University Press, New York, pp 185–195

Bandelt HJ, Dress AWM (1992a) A canonical decomposition-theory for metrics on a finite-set. Adv Math 92:47–105

Bandelt HJ, Dress AWM (1992b) Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol Phylogenet Evol 1:242–252

Billionnet A (2013) Solution of the generalized Noah's Ark problem. Syst Biol 62:147–156

Bordewich M, Semple C (2008) Nature reserve selection problem: a tight approximation algorithm. IEEE/ACM Trans Comput Biol Bioinform 5:275–280

Bordewich M, Rodrigo AG, Semple C (2008) Selecting taxa to save or sequence: desirable criteria and a greedy solution. Syst Biol 57:825–834

Bordewich M, Semple C, Spillner A (2009) Optimizing phylogenetic diversity across two trees. Appl Math Lett 22:638–641

Bottrill MC, Joseph LN, Carwardine J, Bode M, Cook C, Game ET, Grantham H, Kark S, Linke S, McDonald-Madden E, Pressey RL, Walker S, Wilson KA, Possingham HP (2008) Is conservation triage just smart decision making? Trends Ecol Evol 23:649–654

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21:255–265

Chernomor O, Minh BQ, Forest F, Klaere S, Ingram T, Henzinger M, von Haeseler A (2015) Split diversity in constrained conservation prioritization using integer programming. Methods Ecol Evol 6:83–91

Church RL, Stoms DM, Davis FW (1996) Reserve selection as a maximal covering location problem. Biol Conserv 76:105–112

Cocks KD, Baird IA (1989) Using mathematical-programming to address the multiple reserve selection problem – an example from the Eyre Peninsula, South-Australia. Biol Conserv 49:113–130

CPLEX (2012) IBM ILOG CPLEX optimizer

Dantzig G, Fulkerson R, Johnson S (1954) Solution of a large-scale traveling-salesman problem. J Oper Res Soc Am 2:393–410

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61:1–10

Faith DP, Reid CAM, Hunter J (2004) Integrating phylogenetic diversity, complementarity, and endemism for conservation assessment. Conserv Biol 18:255–261

Gomory RE (1958) Outline of an algorithm for integer solutions to linear programs. Bull Am Math Soc 64:275–278

Gurobi Optimization Inc (2013) Gurobi optimizer reference manual

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

Hickey G, Carmi P, Maheshwari A, Zeh N (2008) NAPX: a polynomial time approximation scheme for the Noah's Ark problem. Algoritm Bioinforma Wabi 5251:76–86

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267

Huson DH, Rupp R, Scornavacca C (2010) Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press, Cambridge

Jermiin LS, Jayaswal V, Ababneh F, Robinson J (2008) Phylogenetic model evaluation. In: Keith (ed) Bioinformatics: data, sequences analysis and evolution. Humana Press, Totowa, pp 331–363

Jünger M, Liebling TM, Naddef D, Nemhauser GL, Pulleyblank WR, Reinelt G, Rinaldi G, Wolsey LA (2010) 50 years of integer programming 1958–2008: from the early years to the state-of-the-art. Springer, Heidelberg

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO (2010) Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464

Kimball RT, Braun EL (2008) A multigene phylogeny of Galliformes supports a single origin of erectile ability in non-feathered facial traits. J Avian Biol 39:438–445

Kirkpatrick JB (1983) An iterative method for establishing priorities for the selection of nature reserves: an example from Tasmania. Biol Conserv 25:127–134

May RM (1990) Taxonomy as destiny. Nature 347:129–130

Minh BQ, Klaere S, von Haeseler A (2006) Phylogenetic diversity within seconds. Syst Biol 55:769–773

Minh BQ, Klaere S, von Haeseler A (2009a) Taxon selection under split diversity. Syst Biol 58:586–594

Minh BQ, Pardi F, Klaere S, von Haeseler A (2009b) Budgeted phylogenetic diversity on circular split systems. IEEE/ACM Trans Comput Biol Bioinform 6:22–29

Minh BQ, Klaere S, von Haeseler A (2010) SDA*: a simple and unifying solution to recent bioinformatic challenges for conservation genetics. In: Pham SB, Hoang TH, McKay B, Hirota K (eds) The second international conference on knowledge and systems engineering. IEEE Computer Society, Hanoi, pp 33–37

Minh BQ, Nguyen MAT, von Haeseler A (2013) Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol 30:1188–1195

Moilanen A, Kujala H, Leathwick JR (2009) The zonation framework and software for conservation prioritization. In: Moilanen A, Wilson KA, Possingham HP (eds) Spatial conservation prioritization: quantitative methods and computational tools. Oxford University Press, New York

Moulton V, Semple C, Steel M (2007) Optimizing phylogenetic diversity under constraints. J Theor Biol 246:186–194

Pardi F, Goldman N (2005) Species choice for comparative genomics: being greedy works. PLoS Genet 1:e71

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

Possingham HP, Ball IR, Andelman S (2000) Mathematical methods for identifying representative reserve networks. In: Ferson S, Burgman M (eds) Quantitative methods for conservation biology. Springer, New York, pp 291–305

Pressey RL, Possingham HP, Day JR (1997) Effectiveness of alternative heuristic algorithms for identifying indicative minimum requirements for conservation reserves. Biol Conserv 80:207–219

Rodrigues ASL, Gaston KJ (2002) Maximising phylogenetic diversity in the selection of networks of conservation areas. Biol Conserv 105:103–111

Rodrigues ASL, Brooks TM, Gaston KJ (2005) Integrating phylogenetic diversity in the selection of priority areas for conservation: does it make a difference? In: Purvis A, Gittleman JL, Brooks T (eds) Phylogeny and conservation. Cambridge University Press, Cambridge, pp 101–119

Spillner A, Nguyen BT, Moulton V (2008) Computing phylogenetic diversity for split systems. IEEE/ACM Trans Comput Biol Bioinform 5:235–244

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

Underhill LG (1994) Optimal and suboptimal reserve selection algorithms. Biol Conserv 70:85–87

van der Heide CM, van den Bergh JCJM, van Ierland EC (2005) Extending Weitzman's economic ranking of biodiversity protection: combining ecological and genetic considerations. Ecol Econ 55:218–223

Vane-Wright RI, Humphries CJ, Williams PH (1991) What to protect – systematics and the agony of choice. Biol Conserv 55:235–254

Volkmann L, Martyn I, Moulton V, Spillner A, Mooers AO (2014) Prioritizing populations for conservation using phylogenetic networks. PLoS One 9:e88945

Wang N, Kimball RT, Braun EL, Liang B, Zhang ZW (2013) Assessing phylogenetic relationships among Galliformes: a multigene phylogeny with expanded taxon sampling in Phasianidae. PLoS One 8:e64312

Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24:2098–2100

Weitzman ML (1992) On diversity. Q J Econ 107:363–405

Weitzman ML (1998) The Noah's Ark problem. Econometrica 66:1279–1298

Witting L, Loeschcke V (1995) The optimization of biodiversity conservation. Biol Conserv 71:205–207

Witting L, Tomiuk J, Loeschcke V (2000) Modelling the optimal conservation of interacting species. Ecol Model 125:123–143

Wolsey LA (1998) Integer programming. Wiley-Interscience, New York