Submodular Functions and Optimizing Biodiversity

Charles Semple

School of Mathematics and Statistics
University of Canterbury
New Zealand

La Vacquerie 2015
Conservation biology

*Measures* are used in conservation biology to quantify the biological diversity of a collection of species.

These measures are used for selecting which species to conserve.
Conservation biology

Measures are used in conservation biology to quantify the biological diversity of a collection of species.

These measures are used for selecting which species to conserve.

Typically, individual species are the focus of attention, but this is not necessarily the best way to conserve diversity:

*Although conservation action is frequently targeted toward single species, the most effective way of preserving overall species diversity is by conserving viable populations in their natural habitats, often by designating networks of protected areas.* Rodrigues et al. (2005)
Phylogenetic diversity (PD) is a quantitative tool for measuring the biodiversity of a collection of species (Faith 1992).

For a subset $Y$ of species, $PD(Y)$ is the evolutionary distance spanned by the species in $Y$.

- This distance is usually defined in reference to some phylogeny, but here we (initially) extend this with reference to a collection of 2-state characters.
Phylogenetic diversity on splits

A bipartition \(\{A, B\}\) of a set \(X\), where \(|A|, |B| \geq 1\), is a split of \(X\).

A split system \(\Sigma\) of \(X\) is a collection of splits of \(X\).

\(\Sigma\) is weighted if there is a map \(w : \Sigma \to \mathbb{R}^{\geq 0}\).

For a subset \(Z\) of \(X\),

\[
PD(Z) = \sum_{A|B \in \Sigma; A \cap Z, B \cap Z \neq \emptyset} w(A|B).
\]
Phylogenetic diversity on splits

Example $w(\text{ag}|\text{bcdef}) = 1$, $w(\text{abg}|\text{cdef}) = 5$, $w(\text{abefg}|\text{cd}) = 2$, $w(\text{a}|\text{−}) = 1$, $w(\text{b}|\text{−}) = 3$, $w(\text{c}|\text{−}) = 1$, $w(\text{d}|\text{−}) = 2$, $w(\text{e}|\text{−}) = 3$, $w(\text{f}|\text{−}) = 2$, $w(\text{g}|\text{−}) = 2$
Phylogenetic diversity on splits

Example \( w(\text{ag}|\text{bcdef}) = 1, w(\text{abg}|\text{cdef}) = 5, w(\text{abefg}|\text{cd}) = 2, \)
\( w(\text{a}|\cdot) = 1, w(\text{b}|\cdot) = 3, w(\text{c}|\cdot) = 1, w(\text{d}|\cdot) = 2, w(\text{e}|\cdot) = 3, \)
\( w(\text{f}|\cdot) = 2, w(\text{g}|\cdot) = 2 \)

If \( Z = \{a, b, f\} \), then

\[
PD(Z) = ag|bcdef + abg|cdef + a|\cdot + b|\cdot + f|\cdot = 1 + 5 + 1 + 3 + 2 = 12
\]
Phylogenetic diversity on splits

Example \( w(\text{ag}|\text{bcdef}) = 1, \ w(\text{abg}|\text{cdef}) = 5, \ w(\text{abefg}|\text{cd}) = 2, \ w(\text{a}|-) = 1, \ w(\text{b}|-) = 3, \ w(\text{c}|-) = 1, \ w(\text{d}|-) = 2, \ w(\text{e}|-) = 3, \ w(\text{f}|-) = 2, \ w(\text{g}|-) = 2 \)

If \( Z = \{a, b, f\} \), then

\[
PD(Z) = \text{ag} | \text{bcdef} + \text{abg} | \text{cdef} + \text{a}|- + \text{b}|- + \text{f}|- \\
= 1 + 5 + 1 + 3 + 2 = 12
\]
Phylogenetic diversity on splits

Example \( w(\text{ag}|\text{bcdef}) = 1, \ w(\text{abg}|\text{cdef}) = 5, \ w(\text{abefg}|\text{cd}) = 2, \ w(\text{a}|\text{−}) = 1, \ w(\text{b}|\text{−}) = 3, \ w(\text{c}|\text{−}) = 1, \ w(\text{d}|\text{−}) = 2, \ w(\text{e}|\text{−}) = 3, \ w(\text{f}|\text{−}) = 2, \ w(\text{g}|\text{−}) = 2 \)

If \( Z = \{a, b, f\} \), then

\[
PD(Z) = \text{ag|bcdef} + \text{abg|cdef} + \text{a|−} + \text{b|−} + \text{f|−} \\
= 1 + 5 + 1 + 3 + 2 = 12
\]
Phylogenetic diversity on splits

Example \( w(ag|bcdef) = 1, w(abg|cdef) = 5, w(abefg|cd) = 2, \)
\( w(a|\neg) = 1, w(b|\neg) = 3, w(c|\neg) = 1, w(d|\neg) = 2, w(e|\neg) = 3, \)
\( w(f|\neg) = 2, w(g|\neg) = 2 \)

If \( Z = \{a, b, f\} \), then

\[
PD(Z) = ag|bcdef + abg|cdef + a|\neg + b|\neg + f|\neg \\
= 1 + 5 + 1 + 3 + 2 = 12
\]
Phylogenetic diversity across reserves

For a weighted split system $\Sigma$ of $X$ and a collection $\mathcal{R}$ of protected reserves containing species in $X$, the phylogenetic diversity of a subset $S$ of $\mathcal{R}$ is the PD score of the species contained within at least one reserve in $S$. 

**Example**

$$PD(\{\{a, b\}, \{c, e\}, \{a, g, e\}\}) = PD(\{a, b, c, e, g\}) = 18$$
Phylogenetic diversity across reserves

For a weighted split system $\Sigma$ of $X$ and a collection $\mathcal{R}$ of protected reserves containing species in $X$, the phylogenetic diversity of a subset $S$ of $\mathcal{R}$ is the PD score of the species contained within at least one reserve in $S$.

Example $S = \{\{a, b\}, \{c, e\}, \{a, g, e\}\}$

$PD(\{\{a, b\}, \{c, e\}, \{a, g, e\}\}) = PD(\{a, b, c, e, g\}) = 18$
Phylogenetic diversity across reserves

For a weighted split system $\Sigma$ of $X$ and a collection $R$ of protected reserves containing species in $X$, the phylogenetic diversity of a subset $S$ of $R$ is the PD score of the species contained within at least one reserve in $S$.

Example $S = \{\{a, b\}, \{c, e\}, \{a, g, e\}\}$

$PD(\{\{a, b\}, \{c, e\}, \{a, g, e\}\}) = PD(\{a, b, c, e, g\}) = 18$
The problem

**Budgeted Nature Reserve Selection (BNRS)**

*Instance:* A weighted split system $\Sigma$ on $X$, a collection $\mathcal{R}$ of regions containing species in $X$, a cost of preservation for each region, a fixed budget $B$. 

Applications of using phylogenetic diversity across regions to make assessments in conservation planning include Moritz and Faith (1998), Rodrigues and Gaston (2002), Smith et al. (2000).
The problem

**Budgeted Nature Reserve Selection (BNRS)**

**Instance:** A weighted split system $\Sigma$ on $X$, a collection $\mathcal{R}$ of regions containing species in $X$, a cost of preservation for each region, a fixed budget $B$.

**Task:** Find a subset of regions to preserve that maximizes the PD score on $\Sigma$ of the species contained within the preserved regions while keeping within budget $B$. 

Applications of using phylogenetic diversity across regions to make assessments in conservation planning include Moritz and Faith (1998), Rodrigues and Gaston (2002), Smith et al. (2000).
The problem

**Budgeted Nature Reserve Selection (BNRS)**

**Instance:** A weighted split system $\Sigma$ on $X$, a collection $\mathcal{R}$ of regions containing species in $X$, a cost of preservation for each region, a **fixed budget** $B$.

**Task:** Find a subset of regions to preserve that maximizes the PD score on $\Sigma$ of the species contained within the preserved regions while keeping within budget $B$.

Applications of using phylogenetic diversity across regions to make assessments in conservation planning include Moritz and Faith (1998), Rodrigues and Gaston (2002), Smith et al. (2000).
Budgeted Nature Reserve Selection

If $\mathcal{R}$ consists of all singleton subsets of $X$ with each subset having unit cost and

- $\Sigma$ is compatible, then $\text{BNRS}$ is solvable in polynomial time (Faith, 1992; Pardi and Goldman, 2005; Steel, 2005).
- $\Sigma$ is circular, then $\text{BNRS}$ is solvable in polynomial time (Minh et al., 2009).
- $\Sigma$ is affine, then $\text{BNRS}$ is solvable in polynomial time (Spillner et al., 2008).

However, if $\mathcal{R}$ consists of all singleton subsets of $X$ with each subset having unit cost and $\Sigma$ is arbitrary, then $\text{BNRS}$ is NP-hard (Spillner et al., 2008).
Budgeted Nature Reserve Selection

If $\mathcal{R}$ consists of all singleton subsets of $X$ with each subset having unit cost and

- $\Sigma$ is compatible, then $\text{BNRS}$ is solvable in polynomial time (Faith, 1992; Pardi and Goldman, 2005; Steel, 2005).
- $\Sigma$ is circular, then $\text{BNRS}$ is solvable in polynomial time (Minh et al., 2009).
- $\Sigma$ is affine, then $\text{BNRS}$ is solvable in polynomial time (Spillner et al., 2008).

However, if $\mathcal{R}$ consists of all singleton subsets of $X$ with each subset having unit cost and $\Sigma$ is arbitrary, then $\text{BNRS}$ is NP-hard (Spillner et al., 2008).
Budgeted Nature Reserve Selection

If $\Sigma$ is compatible and each region has unit cost, then BNRS is NP-hard (Moulton, S, Steel 2007).
If $\Sigma$ is compatible and each region has unit cost, then \textbf{BNRS} is NP-hard (Moulton, S, Steel 2007).

\textbf{Theorem} (Bordewich, S 2008)

If $\Sigma$ is compatible, then there is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for \textbf{BNRS}. Moreover, for any $\delta > 0$, \textbf{BNRS} cannot be approximated with an approximation ratio of $(1 - \frac{1}{e} + \delta)$ unless P=NP.

- The algorithm returns a feasible solution whose score is at least $(1 - \frac{1}{e})$ (approx. 0.63) times the optimal score.
If $\Sigma$ is compatible and each region has unit cost, then \textbf{BNRS} is NP-hard (Moulton, S, Steel 2007).

**Theorem (Bordewich, S 2008)**

If $\Sigma$ is compatible, then there is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for \textbf{BNRS}. Moreover, for any $\delta > 0$, \textbf{BNRS} cannot be approximated with an approximation ratio of $(1 - \frac{1}{e} + \delta)$ unless $P=NP$.

The restriction to compatible split systems is redundant. In particular, we have

**Theorem (Bordewich, S 2012)**

There is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for \textbf{BNRS}.
Submodular functions

For a set $E$, a function $f: 2^E \to \mathbb{R}$ is submodular if, for all subsets $S, T \subseteq E$,

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T).$$
Submodular functions

For a set $E$, a function $f : 2^E \to \mathbb{R}$ is submodular if, for all subsets $S, T \subseteq E$,

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T).$$

Optimizing Submodular Functions (OSF)

Instance: A non-negative, non-decreasing, submodular function $f$ on $2^E$ which is computable in polynomial time, a cost function on $E$, a fixed budget $B$. 
Submodular functions

For a set $E$, a function $f : 2^E \to \mathbb{R}$ is submodular if, for all subsets $S, T \subseteq E$,

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T).$$

**Optimizing Submodular Functions (OSF)**

Instance: A non-negative, non-decreasing, submodular function $f$ on $2^E$ which is computable in polynomial time, a cost function on $E$, a fixed budget $B$.

Task: Find a subset $S$ of $E$ which maximizes $f$ while keeping within budget $B$. 
Approximation solution to OSF

**APPROXFunction**

1. **Exhaustively** find a feasible solution of size at most two that maximizes $f$. Call the resulting solution $H_1$.

2. For all subsets of $E$ of size three,
   
   (i) **Sequentially** add elements of $E$ that maximize the ratio of incremental $f$ to cost while keeping within budget.
   
   (ii) Do this until no more elements can be added.

   Call the best solution $H_2$.

3. Output $H_1$ or $H_2$ depending on which has the bigger value.

**Theorem (Sviridenko, 2004)**
**APPROXFunction** is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for OSF.
Lemma Let $\Sigma$ be a weighted split system of $X$, let $Q$ be a non-empty subset of $X$, and let $\mathcal{R}$ be a collection of subsets of $X$. Then the function $PD_{(Q, \Sigma)} : 2^\mathcal{R} \rightarrow \mathbb{R}_{\geq 0}$ defined for all subsets $\mathcal{R}'$ of $\mathcal{R}$, by the PD score of $Q \cup \bigcup_{R \in \mathcal{R}'} R$ is a submodular function.
A PD-based submodular function

Lemma Let $\Sigma$ be a weighted split system of $X$, let $Q$ be a non-empty subset of $X$, and let $\mathcal{R}$ be a collection of subsets of $X$. Then the function $PD(Q, \Sigma) : 2^{\mathcal{R}} \rightarrow \mathbb{R}^{\geq 0}$ defined for all subsets $\mathcal{R}'$ of $\mathcal{R}$, by the PD score of $Q \cup \bigcup_{R \in \mathcal{R}'} R$ is a submodular function.

- Choosing $E$, $f$, $c$, and $B$ to be $\mathcal{R} - Q$, $PD(Q, \Sigma)$, $c$, and $B - c(Q)$, APPROXFUNCTION is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for BNRS for when the selected set of reserves includes $Q$. 
A PD-based submodular function

**Lemma** Let $\Sigma$ be a weighted split system of $X$, let $Q$ be a non-empty subset of $X$, and let $\mathcal{R}$ be a collection of subsets of $X$. Then the function $PD(Q,\Sigma) : 2^\mathcal{R} \to \mathbb{R}_{\geq 0}$ defined for all subsets $\mathcal{R}'$ of $\mathcal{R}$, by the PD score of $Q \cup \bigcup_{R \in \mathcal{R}'} R$ is a submodular function.

- Choosing $E$, $f$, $c$, and $B$ to be $\mathcal{R} - Q$, $PD(Q,\Sigma)$, $c$, and $B - c(Q)$, **APPROXFUNCTION** is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for **BNRS** for when the selected set of reserves includes $Q$.

- By running through each possible choice for $Q$, we get a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for **BNRS**.
**BNRS in the rooted setting**

For a rooted phylogenetic $X$-tree $\mathcal{T}$ and a subset $Y$ of $X$, $PD(Y)$ is the sum of the edge weights of the minimal subtree in $\mathcal{T}$ that connects the elements in $Y$ and the root.

**Theorem (Bordewich, S 2008)**

There is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for $rBNRS$. Moreover, for any $\delta > 0$, $rBNRS$ cannot be approximated with an approximation ratio of $(1 - \frac{1}{e} + \delta)$ unless $P=NP$.

Can we extend $rBNRS$ while maintaining the property of a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for solving it?
**BNRS in the rooted setting**

For a rooted phylogenetic $X$-tree $\mathcal{T}$ and a subset $Y$ of $X$, $PD(Y)$ is the sum of the edge weights of the minimal subtree in $\mathcal{T}$ that connects the elements in $Y$ and the root.

**Theorem (Bordewich, S 2008)**
There is a polynomial-time $(1 - \frac{1}{e})$-approximation algorithm for $\text{rBNRS}$. Moreover, for any $\delta > 0$, $\text{rBNRS}$ cannot be approximated with an approximation ratio of $(1 - \frac{1}{e} + \delta)$ unless $P=NP$. 
For a rooted phylogenetic \( X \)-tree \( T \) and a subset \( Y \) of \( X \), \( PD(Y) \) is the sum of the edge weights of the minimal subtree in \( T \) that connects the elements in \( Y \) and the root.

**Theorem (Bordewich, S 2008)**
There is a polynomial-time \( (1 - \frac{1}{e}) \)-approximation algorithm for \( \text{rBNRS} \). Moreover, for any \( \delta > 0 \), \( \text{rBNRS} \) cannot be approximated with an approximation ratio of \( (1 - \frac{1}{e} + \delta) \) unless \( P=NP \).

Can we extend \( \text{rBNRS} \) while maintaining the property of a polynomial-time \( (1 - \frac{1}{e}) \)-approximation algorithm for solving it?
Extensions

1. Evolutionary relationships are not necessarily represented by a single tree. For example, a collection of gene trees may be a better representation.
   - Allow for a collection of trees.
Evolutionary relationships are not necessarily represented by a single tree. For example, a collection of gene trees may be a better representation.

- Allow for a collection of trees.

It’s unrealistic to expect a species probability of survival is zero if it is not contained in a selected region or, if it is contained in such a region, its probability of survival is one.

- Allow for arbitrary survival probabilities.
Evolutionary relationships are not necessarily represented by a single tree. For example, a collection of gene trees may be a better representation.

- Allow for a collection of trees.

It’s unrealistic to expect a species probability of survival is zero if it is not contained in a selected region or, if it is contained in such a region, its probability of survival is one.

- Allow for arbitrary survival probabilities.

PD assumes that features arise uniformly across a phylogeny and persist to be present in all descendant species. But features may disappear over time.

- Allow for features to disappear.
  Once a feature is present, it has a constant and memoryless probability $e^{-\lambda}$ of surviving in each time step.
PD in the extended rooted setting

Each $x \in X$ has a probability $p(x)$ of survival. Under the extended model, the PD of $X$ on $T$ is the expected number of features present amongst the surviving taxa.
PD in the extended rooted setting

Each \( x \in X \) has a probability \( p(x) \) of survival. Under the extended model, the PD of \( X \) on \( T \) is the expected number of features present amongst the surviving taxa.

\[
PD_{(\lambda, T)}(X, p) = \int_{t \in T} P(t \rightarrow X) dt,
\]

where \( (t \rightarrow X) \) denotes the event that a feature arising at point \( t \) on \( T \) survives to be present in a taxa in \( X \) which itself survives.
PD in the extended rooted setting

Each $x \in X$ has a probability $p(x)$ of survival. Under the extended model, the PD of $X$ on $\mathcal{T}$ is the expected number of features present amongst the surviving taxa.

$$PD_{(\lambda, \mathcal{T})}(X, p) = \int_{t \in \mathcal{T}} \mathbb{P}(t \to X) dt,$$

where $(t \to X)$ denotes the event that a feature arising at point $t$ on $\mathcal{T}$ survives to be present in a taxa in $X$ which itself survives.

Summing over a collection $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_k\}$ of weighted trees, the PD of $X$ on $\mathcal{P}$ is

$$PD_{(\lambda, \mathcal{P})}(X, p) = \sum_{j=1}^{k} w(\mathcal{T}_j) \int_{t \in \mathcal{T}_j} \mathbb{P}(t \to X) dt.$$
rBNRS under the extended model

$\text{BNRS}_{(\lambda, P)}$

**Instance:** A collection $P$ of weighted trees on $X$, a collection $R$ of regions containing species in $X$, a cost of preservation for each region, a fixed budget $B$ and, for all $(x, R) \in X \times R$, probabilities $a(x, R)$ and $b(x, R)$, where $b(x, R) \geq a(x, R)$. 
**rBNRS under the extended model**

\[ \text{BNRS}_{(\lambda, \mathcal{P})} \]

**Instance:** A collection \( \mathcal{P} \) of weighted trees on \( X \), a collection \( \mathcal{R} \) of regions containing species in \( X \), a cost of preservation for each region, a fixed budget \( B \) and, for all \((x, R) \in X \times \mathcal{R}\), probabilities \( a(x, R) \) and \( b(x, R) \), where \( b(x, R) \geq a(x, R) \).

**Task:** Find a subset \( \mathcal{R}' \) of regions to preserve that maximizes \( PD_{(\lambda, \mathcal{P})}(X, p_{\mathcal{R}'}) \) while keeping within budget \( B \).
**rBNRS** under the extended model

**BNRS}_{\lambda,P}\)

**Instance:** A collection $P$ of weighted trees on $X$, a collection $R$ of regions containing species in $X$, a cost of preservation for each region, a fixed budget $B$ and, for all $(x, R) \in X \times R$, probabilities $a(x, R)$ and $b(x, R)$, where $b(x, R) \geq a(x, R)$.

**Task:** Find a subset $\mathcal{R}'$ of regions to preserve that maximizes $PD_{\lambda,P}(X, \rho_{\mathcal{R}'})$ while keeping within budget $B$.

**Lemma** The function $PD_{\lambda,P} : 2^R \rightarrow \mathbb{R}^{\geq 0}$ is non-negative, non-decreasing, submodular, and computable in polynomial time.
rBNRS under the extended model

**BNRS\(_{(\lambda, \mathcal{P})}\)**

**Instance:** A collection \(\mathcal{P}\) of weighted trees on \(X\), a collection \(\mathcal{R}\) of regions containing species in \(X\), a cost of preservation for each region, a fixed budget \(B\) and, for all \((x, R) \in X \times \mathcal{R}\), probabilities \(a(x, R)\) and \(b(x, R)\), where \(b(x, R) \geq a(x, R)\).

**Task:** Find a subset \(\mathcal{R}'\) of regions to preserve that maximizes \(PD_{(\lambda, \mathcal{P})}(X, p_{\mathcal{R}'})\) while keeping within budget \(B\).

**Theorem (Bordewich, S 2012)**

There is a polynomial-time \((1 - \frac{1}{e})\)-approximation algorithm for BNRS\(_{(\lambda, \mathcal{P})}\). Moreover, for any \(\delta > 0\), BNRS\(_{(\lambda, \mathcal{P})}\) cannot be approximated with an approximation ratio of \((1 - \frac{1}{e} + \delta)\) unless P=NP.