Phylogenetic Diversity on Split Systems

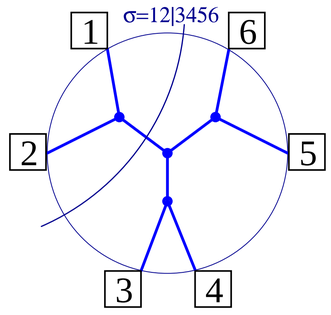
Bui Quang Minh, Steffen Klaere and Arndt von Haeseler
Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories,
University of Vienna, Medical University of Vienna, Veterinary University of Vienna

Phylogenetic Diversity on Trees

Biodiversity: is the variability among living organisms from all sources, including, 'inter alia', terrestrial, marine, and other aquatic ecosystems, and the ecological complexes of which they are part: this includes diversity within species, between species and of ecosystems [UN Earth Summit, 1992].

Phylogenetic Diversity: a measure of the content of feature diversity of a taxon subset $W \subseteq \mathcal{X}$ relative to the entire variation of the phylogenetic tree (sensu Faith, 1992).

Task: Assume that you can only support k taxa. Which subset of k taxa has the highest phylogenetic diversity?

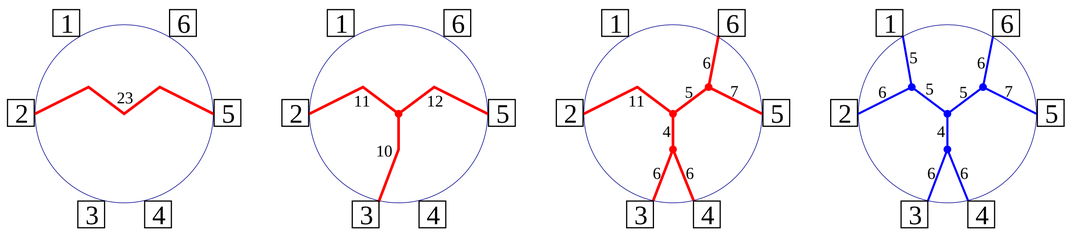


Let $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ denote the tree describing the relationship of the taxon set $\mathcal{X}$ with respect to a certain feature (morphological, gene-level). The nonnegative number $\lambda(e)$ is the length of edge $e \in \mathcal{E}$. The tree $\mathcal{T}_W$ for a taxon subset W contains all connecting edges from $\mathcal{T}$ and the interior nodes whose degree remains larger than two. If an interior node vanishes and its incident edges merge, the length $\lambda_W(.)$ of the new edge is the sum of the length of the merged edges. The diversity of the taxon subset W is the sum of the length of the edges in $\mathcal{E}_W$, i.e.

$$\mathrm{PD}(W) = \sum_{e \in \mathcal{E}_W} \lambda_W(e).$$

Greedy works on Trees

For all taxon sets $W \in \mathrm{PD}_j$ exists a taxon $w \in \mathcal{X} - W$ such that $W \cup \{w\} \in \mathrm{PD}_{j+1}$.



Optimality of greedy algorithm on trees proven independently by Steel (2005) and Pardi and Goldman (2005); fast implementation by Minh *et al.* (2006).

Phylogenetic Diversity on Circular Networks

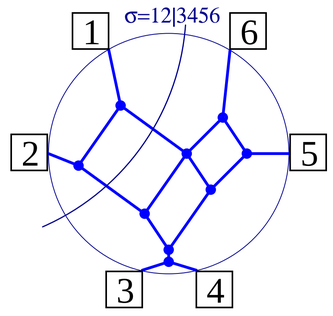
Assume we wish to put our decision not on one but on multiple features. Each feature suggests a tree, and these trees are not necessarily compatible. Which structure can simultaneously regard conflicting information?

Split: Decomposition $A|B$ of taxon set $\mathcal{X}$, i.e. $A \cap B = \emptyset,\ A \cup B = \mathcal{X}$.

Trees: Each edge in a tree uniquely identifies a split. Collecting all splits from the feature trees gives us a split system that might not identify a tree.

Split Networks: This visualization is introduced in Bandelt and Dress (1992). Splits are now illustrated by parallel or single edges in the graph.

Circular: A split system has a circular representation if all taxa can be placed on an circle and each split can be illustrated by a line intersecting the circle exactly twice.

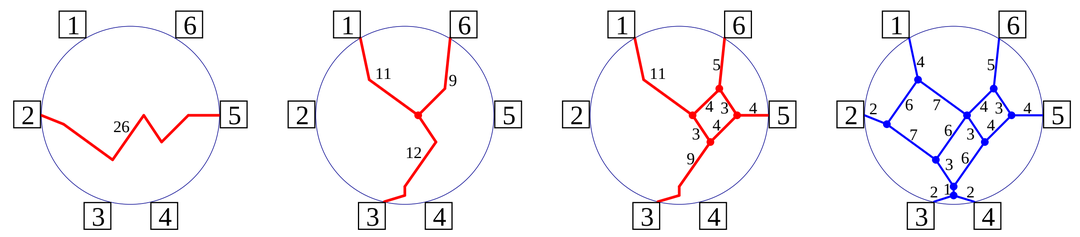


Let Σ denote a split system for taxon set $\mathcal{X}$. For each split $\sigma \in \Sigma$ the symbol $\lambda(\sigma)$ may denote its weight. For a taxon subset W the split subsystem Σ_W is given by $\{A|B \in \Sigma :, \exists a, b \in W : a \in A, b \in B\}$. The phylogenetic diversity of W is then

$$\mathrm{PD}(W) = \sum_{\sigma \in \Sigma_W} \lambda(\sigma).$$

Greedy does not work on Split Systems

Greedy never gives up what it already has acquired. As the example below shows, this is not always a good strategy, especially if the two elements from the optimal 2-set are not in the optimal 3-set.



Therefore, an alternative approach to acquire optimal subsets is needed. And we have a suggestion.

Circular Tours and Phylogenetic Diversity

Define the distance between two taxa u and v as the sum of the weights of splits separating u from v, i.e.

$$d_{uv} = \sum_{\substack{A|B \in \Sigma \\ u \in A, B \in v}} \lambda(A|B).$$

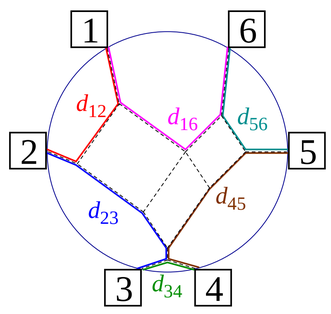
A circular split network always comes with a circular order of the taxa. From results from the traveling salesman problem (e.g., Korostensky and Gonnet, 2000) we learn that for a circular network with circular order $(u_1, u_2, \ldots, u_n)$ of taxa the shortest tour visiting all taxa and returning to the starting taxa is exactly the sum of the distance of consecutive taxa, i.e.

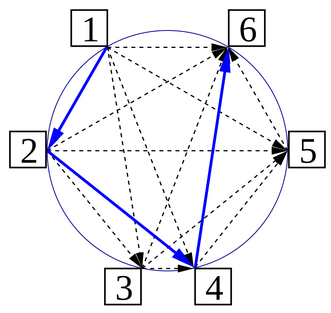
$$\Lambda_{[1,n]} = d_{u_1 u_n} + \sum_{j=1}^{n-1} d_{u_j u_{j+1}},$$

which is also twice the length of the network. Circular order and pairwise distances are retained in subsystems. Therefore, for taxon subset W of size k with inherited taxon order $w_1, \ldots, w_k$ the phylogenetic diversity is calculated by

$$\mathrm{PD}(W) = \frac{1}{2}\left(d_{w_1 w_k} + \sum_{j=1}^{k-1} d_{w_j w_{j+1}}\right).$$

With this the structure on which to optimize can be depicted as an acyclic graph with taxon set $\mathcal{X}$ where each pair (u, v) of taxa is connected by a directed edge with length d_{uv}.





A Dynamic Programming Algorithm

Input: Circular order of taxa $(1, 2, \ldots, n)$, split-distance matrix (d_{uv}), set size k.

Initialization: Length of longest ordered 2-path between taxa $u < v$: $L^2_{uv} = d_{uv}$.

Iteration: In each step i compute for all pairs of taxa $u < v$ the *longest ordered i-path* by

$$L^i_{uv} = \max_{u<w<v} \{L^{i-1}_{uw} + d_{wv}\}.$$

and store L^i_{uv} and $\alpha^i_{uv} = \mathrm{argmax}\, L^i_{uv}$.

Termination: Determine the *longest circular k-tour* by

$$\ell^k = \max_{u,v} \{L^k_{uv} + d_{uv}\}.$$

and the taxa of the k-tour by backtracking through (α^i_{uv}), $i = 2, \ldots, k$.

Optimal Substructure: If $(s_1, s_2, \ldots, s_j)$ is the longest ordered j-path from s_1 to s_j then $(s_1, s_2, \ldots, s_{j-1})$ is the longest ordered $j-1$-path from s_1 to s_{j-1}.

Complexity: $O(kn^3)$ time complexity and $O(kn)$ memory complexity.

Budget considerations

Just fixing the size of the taxon subset is not a very justifiable criterion. Moulton *et al.* (2006) suggest a variety of constraints which will influence the actual selection and give the set of k taxa more biological relevance. Weitzman (1998) proposed an economical side condition, namely that the taxon selection is not fixed by size but by a budget B for supporting the conservation efforts. To this end, each taxon $u \in \mathcal{X}$ is assigned the nonnegative preservation cost $c(u)$. The preservation cost $c(W)$ of a taxon subset W is the sum of preservation costs of its taxa.

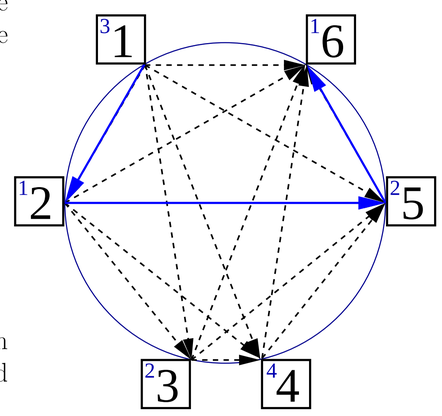
Task: Find the most diverse set W whose preservation cost $c(W)$ does not exceed a given budget B.

Replace in the algorithm the counters from k to B, i.e. regard the symbols L^b_{uv} and α^b_{uv} for $b = 0, 1, \ldots, B$. This replacements retains all properties of the algorithm.

An Example

Reconsider the network introduced in the above panel and assume an overall budget of $B = 7$. The example network introduces the following distance matrix and DAG.

$$(d_{uv}) = \begin{pmatrix} - & 12 & 23 & 23 & 22 & 20 \\ 12 & - & 15 & 15 & 26 & 24 \\ 23 & 15 & - & 4 & 17 & 21 \\ 23 & 15 & 4 & - & 17 & 21 \\ 22 & 26 & 17 & 17 & - & 12 \\ 20 & 24 & 21 & 21 & 12 & - \end{pmatrix}$$



The below table summarizes the optimization process. As we can see, the longest 7-tour has length 70 and starts in taxon 1 and returns to 1 from 6. The associated taxon subset is $\{1, 2, 5, 6\}$.

L^b_{uv}	v	0	1	2	3	4	5	6	7	$L^7_{uv} + d_{uv}$
L^b_{1v}	2	-	-	-	12	12	12	12	12	24
	3	-	-	-	-	-	23	27 (2)	27 (2)	50
	4	-	-	-	-	-	-	-	23	46
	5	-	-	-	-	-	22	38 (2)	40 (3)	62
	6	-	-	-	-	20	36 (2)	44 (3)	**50** (5)	**70**
L^b_{2v}	3	-	-	-	15	15	15	15	15	30
	4	-	-	-	-	-	15	15	19 (3)	34
	5	-	-	-	26	26	32 (3)	32 (3)	32 (4)	58
	6	-	-	21	21	38 (5)	38 (5)	44 (5)	44 (5)	68
L^b_{3v}	4	-	-	-	-	-	-	4	4	8
	5	-	-	-	-	17	17	17	17	34
	6	-	-	-	21	21	29 (5)	29 (5)	29 (5)	50
L^b_{4v}	5	-	-	-	-	-	-	17	17	34
	6	-	-	-	-	-	21	21	29 (5)	50
L^b_{5v}	6	-	-	-	12	12	12	12	12	24

References

Bandelt, H.-J. and Dress, A. W. M. (1992) Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.*, **1**, 242–252.

Faith, D. P. (1992) Conservation Evaluation and Phylogenetic Diversity. *Biol. Conserv.*, **61**, 1–10.

Korostensky, C. and Gonnet, G. H. (2000) Using traveling salesman problem algorithms for evolutionary tree reconstruction. *Bioinformatics*, **16**, 619–627.

Minh, B. Q., Klaere, S. and von Haeseler, A. (2006) Phylogenetic diversity within seconds. *Syst. Biol.*, **55**, 769–773.

Moulton, V., Semple, C. and Steel, M. (2006) Optimizing phylogenetic diversity under constraints. *Journal of Theoretical Biology*, in Press.

Pardi, F. and Goldman, N. (2005) Species choice for comparative genomics: Being greedy works. *PLoS Genet.*, **1**, 672–675.

Steel, M. (2005) Phylogenetic Diversity and the Greedy Algorithm. *Syst. Biol.*, **54**, 527–529.

Weitzman, M. L. (1998) The Noah's Ark problem. *Econometrica*, **66**, 1279–1298.