Submitted: December 14, 2007

Phylogenetic Diversity on Split Networks

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

INI Preprint Number: N107090

Abstract

In biodiversity conservation, one is interested in selecting a subset of taxa for preservation priority. Phylogenetic diversity (*PD*) provides a quantitative measure for taxon selection on phylogenetic trees. In particular, *PD* is the total length of the minimal subtree induced by the selected taxa. Recently, it has been shown that on trees the maximal *PD* score and the corresponding subset of taxa can be computed by a greedy algorithm. However, if evolution is not treelike and networks are a more appropriate illustration of phylogenetic relationships, then the greedy strategy no longer works.

Here, we will extend the notion of *PD* to phylogenetic networks. To this end, we suggest a dynamic programming algorithm (`PD-NET`) which guarantees the computation of optimal *PD* scores and *PD* sets for circular networks, a commonly encountered category of networks. `PD-NET` has polynomial time complexity. Finally we apply `PD-NET` to biological data and compare the resulting *PD* sets to the selection of taxa derived from a tree. The outcome indicates that it is advisable to include also non-treelike effects when dealing with conservation questions.

Keywords: phylogenetic diversity, dynamic programming, phylogenetic network, split system, biodiversity conservation.

Introduction

Biodiversity embraces the variety of life from plants to animals, from micro- to macroorganisms, from genes to genomes and ecosystems. The conservation planning of biodiversity is concerned with many research projects and intense discussions (e.g., Wilson, 1997; Gaston and Spicer, 2004).

In the last decades, the diversity of a set of taxa has been primarily measured by genetic distance (Vane-Wright et al., 1991), i.e. by the discrepancy between the genetic information of taxa. In particular, one is interested in selecting a subset of k representative taxa which maximize the total genetic distance of all evolutionary lineages spanned by these taxa. This concept was further extended to comparative genomics in prioritizing taxa under sequencing projects (Pardi and Goldman, 2005).

Which measure of genetic distance to use is the subject of numerous discussions (e.g., Nee and May, 1997; Crozier et al., 2005; Faith and Baker, 2006). So far, two measures are primarily used to evaluate the diversity of a taxon set. Genetic diversity is the probability of observing at least two alleles in a given taxon set (Crozier, 1992), whereas feature diversity is the average number of substitutions necessary to observe the considered taxa (Faith, 1992). However, we will not delve into this discussion. The scope of the paper is rather a distance based approach to detect a collection of k representative taxa if the relationship of the group under study is not tree-like.

The evolutionary history is usually assumed to be treelike and the diversity of a set of taxa equals the length of the minimal subtree connecting them. Under these assumptions, Steel (2005) and Pardi and Goldman (2005) have shown that a greedy algorithm is sufficient to determine an optimal taxon set of a given size. Minh et al. (2006) presented an efficient implementation capable of handling trees with thousands of taxa or more.

However, it is well known that different regions of the genome provide trees with different genetic distances between taxa due to violations of the molecular clock or due to varying rates of molecular evolution (e.g., Graur and Li, 2000). Moreover, sometimes different regions of the genome lead to distinct trees due to ancestral polymorphisms (e.g., Nei, 1987). Recently, it has been shown that the most likely gene tree, which is often considered as the species tree may in fact differ from the true species tree (Degnan and Rosenberg, 2006). Furthermore, horizontal gene transfer, frequently exploited among bacteria, is a mechanism leading to non-treelike evolution (Doolittle et al., 2003). Hybridization and recombination are additional factors contributing to inadequate representation of phylogenetic relationships by a single tree.

Therefore, basing a conservation decision on a single tree which cannot depict conflicting signals appears to be rather unjustifiable. A possible solution for this problem is to employ split networks (Huber and Moulton, 2005, and references therein). Such networks are a generalization of phylogenetic trees and allow the user to visualize contradictory signals in data, which cannot be included in trees. A split network represents a set of so-called “splits”, i.e. bipartitions of the taxon set (e.g., Bandelt and Dress, 1992; Dress et al., 2001).

Split networks have been regularly employed in the analyses of bacterial and plant sequences or their allelic diversity (e.g., Sullivan et al., 2006; Henz et al., 2005; Suerbaum et al., 2001; de las Rivas et al., 2004; Hertel et al., 2006). If one is interested in maintaining diversity, it is thus in some instances more appropriate to measure genetic distance on split networks rather than on trees. Unfortunately, the greedy strategy is not applicable to networks (Minh et al., 2006). Therefore an alternative strategy is needed. Such a strategy is the main subject of this paper.

First, we extend the notion of genetic distance to split networks. Then, we introduce an efficient algorithm (`PD-NET`) to obtain an optimal taxon set for so-called circular networks (Bandelt and Dress, 1992), a commonly encountered category of split networks. Such a graph can be reconstructed by e.g. *Neighbor-net* (Bryant, 2004; Huson and Bryant, 2006). `PD-NET` works for unrooted networks and slightly modified for split networks directed by an outgroup (e.g., Faith and Baker, 2006, for a discussion). Subsequently, we conduct a comparative analysis on several real world data sets to show how the optimal taxon sets change when inferred from the split network rather than from the tree derived from the same data set. The outcome indicates that the split network significantly influences the resulting taxon sets, thus exhibiting a promising perspective of the method for real-world applications.

Split Systems and Phylogenetic Diversity

Let $X = \{s_1, \ldots, s_n\}$ denote a finite set of n taxa. A *split* $A|B$ is a bipartition of X into two non-empty disjoint sets A and B, i.e., $A \cap B = \emptyset$ and $A \cup B = X$. A *split system* Σ is any collection of splits of X.

Split systems are usually visualized as *split networks*, where each split is represented by one or several parallel edges (e.g., Huber and Moulton, 2005). Figure 1 displays a five-taxon split system consisting of nine splits and its corresponding split network. As an example for the visualization of a split consider the dashed lines. Removing them from the network would separate the taxon set $\{1, 2\}$ from $\{3, 4, 5\}$. Trees are a special case of networks where each split is represented by exactly one edge, i.e., only *compatible* splits are allowed. For a more technical definition see Bandelt and Dress (1992).

We are particularly interested in *circular* split systems, i.e., a split system Σ for which a *circular taxon order* $(s_1, s_2, \ldots, s_n)$ exists such that all elements of Σ are of the form $\{s_i, s_{i+1}, \ldots, s_j\} \mid X - \{s_i, s_{i+1}, \ldots, s_j\}$, $1 \leq i \leq j \leq n$ (e.g., Bandelt and Dress, 1992). Such split systems can be represented by so-called *outer-labelled plane splits graphs* (Dress and Huson, 2004). Looking at its network representation, the taxa of a circular split network can be placed on a circle and each split can be depicted as a line bisecting the circle. The network in Figure 1 has the circular order $(1, 2, 3, 4, 5)$.

An unrooted tree contains at most $2n-3$ splits which equals the number of branches in a bifurcating unrooted tree, whereas the maximum number of splits in a circular split system is $\binom{n}{2}$ (e.g., Bandelt and Dress, 1992).

We now come to the definition of diversity on networks. Since we calculate the diversity in a phylogenetic framework, we will call the score of a taxon set its *phylogenetic diversity* (*PD*). This term is often used for Faith's feature diversity (e.g., Moulton et al., 2007) but the reader should be aware that the measure is defined for arbitrary functions of (genetic) distance.

Let λ denote the split weight function that assigns to each split $A|B \in \Sigma$ a non-negative weight $\lambda(A|B)$. For any two taxa u and v the pairwise split-distance d_{uv} is the sum of the weights of all splits separating u from v, i.e.,

$$d_{uv} = \sum_{\substack{A|B \in \Sigma \\ u \in A, v \in B \text{ or} \\ v \in A, u \in B}} \lambda(A|B). \qquad (1)$$

As on trees the pairwise distance d_{uv} is also the phylogenetic diversity of the taxon set $\{u, v\}$, i.e. $PD(\{u, v\}) = d_{uv}$. We extend the definition to taxon subsets $S \subset X$ in the following way. The *phylogenetic diversity* of S is the sum of the weights of all splits still separating the taxa of S, i.e.,

$$PD(S) = \sum_{\substack{A|B\in\Sigma \\ A\cap S\neq\emptyset \\ B\cap S\neq\emptyset}} \lambda(A|B). \tag{2}$$

This definition consistently extends the tree-based *PD* (see also Moulton et al., 2007) since splits define the edges of trees.

Note that the *PD* strongly depends on the definition of the split weight function λ. One can propose the split weights to denote the number of substitutions observed between the split sets or declare alternative measures (e.g., Crozier, 1992). However, our aim is not to measure diversity in itself but rather to present a tool that computes diversity for subsets based on arbitrary split weight functions.

Based on split-distances, a key property of circular split systems is that for any subset $S \subset X$ containing k taxa with circular order $(s_1, s_2, \ldots, s_k)$ the *PD* score of S can be computed employing a *circular tour* (Korostensky and Gonnet, 2000). A circular tour visits all taxa and returns to its starting taxon while taking the shortest path connecting s_i and s_{i+1} in the split system. Since each split bisects the circle, a circular tour traverses each split exactly twice. Thus the sum of the weights of all edges encountered during a circular tour equals twice the sum of the weight of all splits. Since circularity is retained for subsystems of circular split systems, the *PD* score of the taxon subset S with circular order $(s_1, s_2, \ldots, s_k)$ is given by

$$PD(S) = \frac{1}{2}\left(d_{s_1 s_k} + \sum_{i=1}^{k-1} d_{s_i s_{i+1}}\right). \tag{3}$$

Thus obtaining the optimal *PD* set for k taxa is equivalent to determining the longest circular tour traversing k taxa.

We introduce a few terms. For any given $2 \leq k \leq n-1$ the maximal *PD* is denoted by $pd_{\max}(k)$, and PD_k denotes the set of all taxon sets S with $PD(S) = pd_{\max}(k)$.

On trees one employs a greedy strategy to obtain the optimal *PD* score and *PD* sets (e.g., Pardi and Goldman, 2005) for a given size k. To this end, one simply constructs an optimal $S \in PD_k$ set by determining an optimal set of two taxa and subsequently adding $k-2$ taxa. Therefore, we have $PD_2 \subset PD_3 \subset \cdots \subset PD_{n-1}$, i.e. for any optimal taxon set $S \in PD_k$ exists a series of taxon sets $(S_j)_{j=2}^{n-1}$ with $S_i \in PD_i$, $S_k = S$ and $S_i \subset S_j$, $i \leq j$. However, when looking at the split system depicted in Figure 1 we find the single optimal PD_2 set is $\{2, 5\}$ and the optimal PD_3 set is $\{1, 3, 4\}$ which are disjoint. Thus the greedy strategy does not apply on general split systems (see also Minh et al., 2006).

PD-NET: An efficient algorithm to obtain optimal *PD* sets for circular split systems

We introduce an efficient algorithm to select k taxa which maximize the *PD* score over all possible sets of k taxa in a circular split system Σ. Equation (3) permits a direct computation of the PD score from a split-distance matrix without considering the detailed structure of the underlying split network. Based on this observation, the calculation of optimal PD sets reduces to the following task:

> *Given n taxa, their circular order and pairwise split-distances (d_{uv}), find the* **longest circular k-tour***, i.e. the longest among those circular tours traversing k taxa.*

Without loss of generality, we assume that $(1, 2, \ldots, n)$ represents the circular order. We construct a directed acyclic graph (DAG; Cormen et al., 2001) with n taxa with respect to the above circular order by introducing a directed edge from u to v with edge length d_{uv} if u precedes v in the circular order. Figure 2 depicts the DAG constructed from the example split network of five taxa with circular order $(1, 2, 3, 4, 5)$.

A collection of k taxa $(s_1, s_2, \ldots, s_k)$ is called an *ordered k-path* if $1 \leq s_1 < s_2 < \cdots < s_k \leq n$. Its length is given by $L(s_1, s_2, \ldots, s_k) = \sum_{i=1}^{k-1} d_{s_i s_{i+1}}$. A circular k-tour is attained if we add the vertex $s_{k+1} = s_1$, i.e., returning to the starting taxon. Clearly, the length of this circular k-tour is then $L(s_1, s_2, \ldots, s_k, s_1) = L(s_1, s_2, \ldots, s_k) + d_{s_1 s_k}$. For two taxa $u < v$ we denote by L^k_{uv} the length of the longest ordered k-path from u to v, i.e.,

$$L^k_{uv} = \max_{u < s_2 < \cdots < s_{k-1} < v} L(u, s_2, \ldots, s_{k-1}, v).$$

It is worth noting that every circular k-tour is uniquely represented by an ordered k-path and vice versa. Therefore, looking at all ordered paths is sufficient to find a longest circular k-tour. We will now present a way to obtain a PD_k set by computing L^i_{uv} for all i from 2 to k, i.e., the length of the longest ordered i-path between every pair of taxa $u < v$.

The key property we make use of is that if $(s_1, s_2, \ldots, s_k, s_1)$ is the longest circular k-tour then $(s_1, s_2, \ldots, s_k)$ is the longest ordered k-path from s_1 to s_k. The proof is given in the appendix.

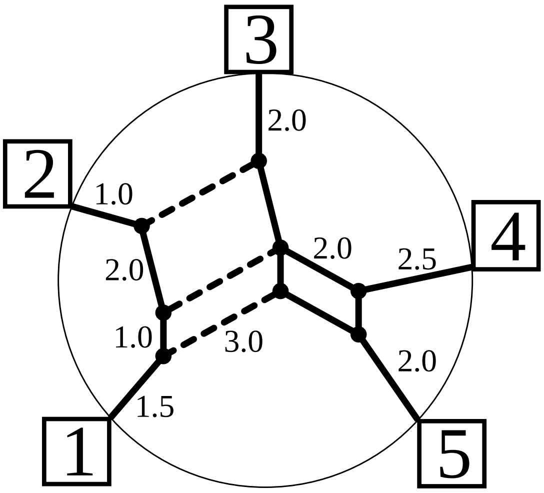
As a result, the length ℓ^k_{max} of the longest circular k-tour will be obtained by solving the following iterative maximization:

$$\ell^k_{max} = \max_{1 \leq u < v \leq n} \{L^k_{uv} + d_{uv}\}, \tag{4}$$

$$L^i_{uv} = \begin{cases} \max_{u<s<v}\{L^{i-1}_{us} + d_{sv}\}, & \text{if } 3 \leq i \leq k, \\ d_{uv}, & \text{if } i = 2. \end{cases} \tag{5}$$

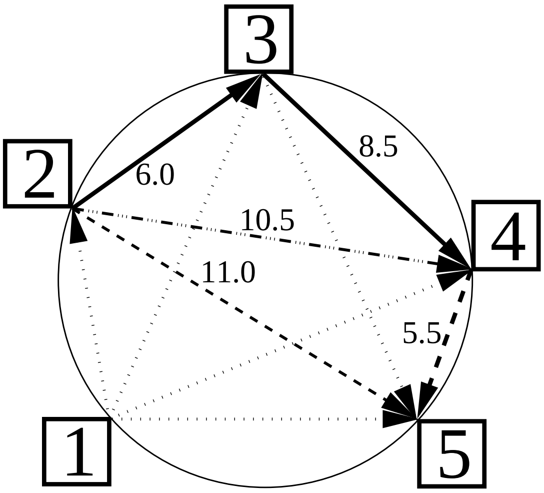
We solve this series of equations by employing a dynamic programming technique in a bottom-up fashion: First compute $L^2_{uv}, L^3_{uv}, \ldots, L^k_{uv}$ for all pairs of taxa $u < v$ by equation (5), and then calculate ℓ^k_{max} by equation (4).

Now based on equation (3), the optimal k-sets $S \in PD_k$ have score $pd_{\max}(k) = \ell^k_{max}/2$. To construct a set S with optimal score, we trace back the taxa which maximize the sum on the right-hand side of the equations (4) and (5).



Split	Weight
1|2345	1.5
2|1345	1.0
3|1245	2.0
4|1235	2.5
5|1234	2.0
12|345	3.0
23|145	2.0
45|123	2.0
15|234	1.0

Figure 1: A sample split system and its corresponding split network of five taxa. A split is depicted by a single or parallel lines. E.g., the split 12|345 is depicted by the dashed lines. The circle connecting the taxa of the graph indicates the circular order.



3-path	path	tour
	length	length
(123)	**11.5**	**21**
(124)	16.0	26
(134)	**18.0**	**28**
(125)	16.5	25
(135)	**18.5**	**27**
(145)	15.5	24
(234)	**14.5**	**25**
(235)	15.0	26
(245)	**16.0**	**27**
(345)	**14.0**	**23**

Figure 2: The directed acyclic graph (DAG) of k-paths. The solid arcs depict the 3-path $(2, 3, 4)$. The table on the right hand side gives the length of all 3-paths and the length of the ten circular tours.

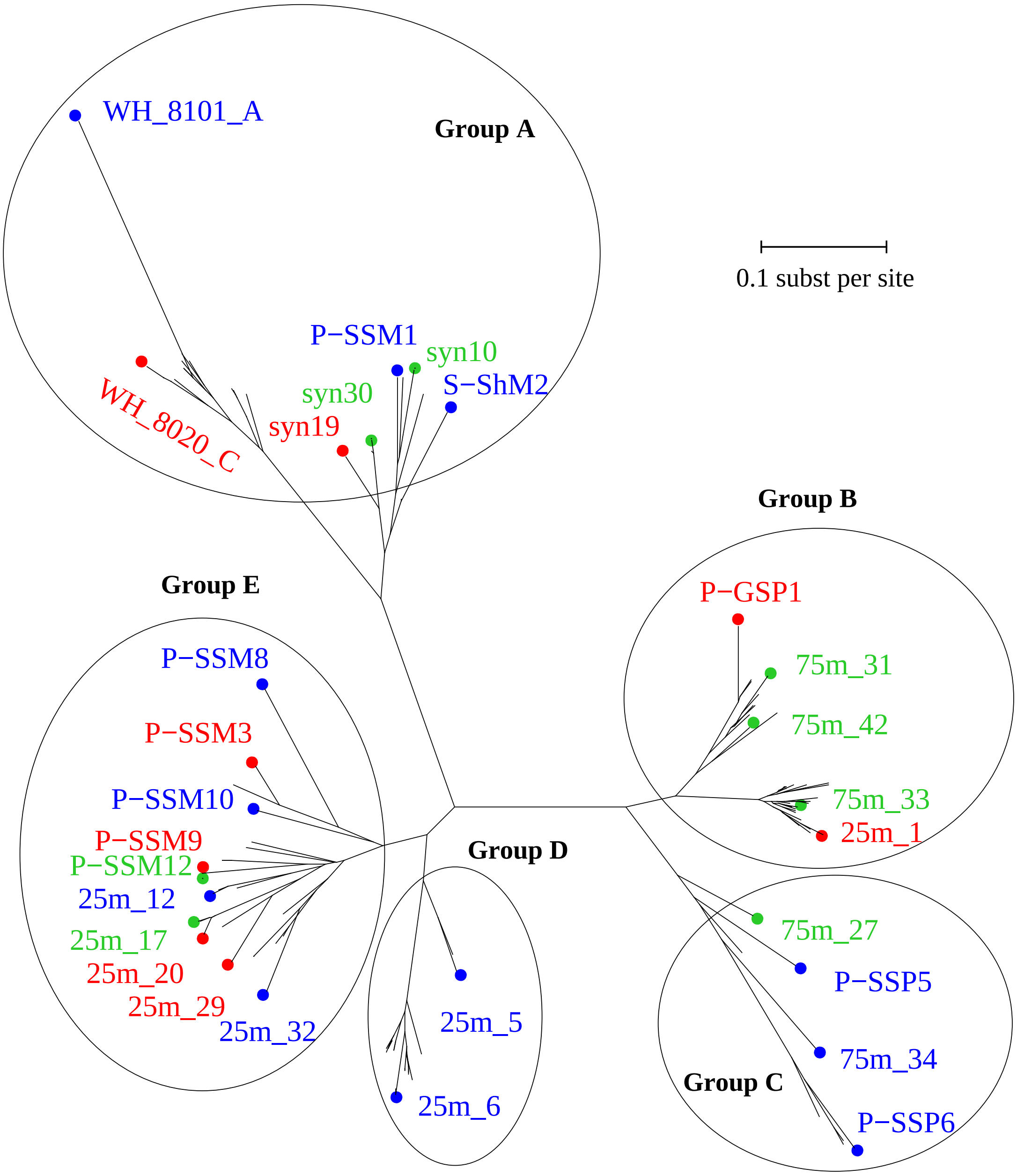


Figure 3: The neighbor-joining tree of the CYANO data with taxa in the union of PD_{20}^{NJ} and PD_{20}^{NNet}. The blue taxa appear in both sets. The red taxa occur exclusively in PD_{20}^{NJ}. The green taxa are in PD_{20}^{NNet} and not in PD_{20}^{NJ}.

An Example

For illustration, let us regard the circular split system in Figure 1. We will construct an optimal PD_3 set. The circular order of the five taxa is $(1, 2, 3, 4, 5)$. Equation (1) leads to the pairwise split-distance matrix:

$$(d_{uv}) = \begin{pmatrix} 0 & 5.5 & 9.5 & 10.0 & 8.5 \\ 5.5 & 0 & 6.0 & 10.5 & 11.0 \\ 9.5 & 6.0 & 0 & 8.5 & 9.0 \\ 10.0 & 10.5 & 8.5 & 0 & 5.5 \\ 8.5 & 11.0 & 9.0 & 5.5 & 0 \end{pmatrix}$$

As noted the length of the longest ordered 2-path L^2_{uv} equals d_{uv} and therefore:

$$(L^2_{uv}) = \begin{pmatrix} - & 5.5 & 9.5 & 10.0 & 8.5 \\ & - & 6.0 & 10.5 & 11.0 \\ & & - & 8.5 & 9.0 \\ & & & - & 5.5 \\ & & & & - \end{pmatrix}$$

From L^2_{uv} we derive L^3_{uv} as described in equation (5):

$$(L^3_{uv}) = \begin{pmatrix} - & - & 11.5 & 18.0 & 18.5 \\ & - & - & 14.5 & 16.0 \\ & & - & - & 14.0 \\ & & & - & - \\ & & & & - \end{pmatrix}$$

where the secondary diagonal elements are omitted since there is no ordered 3-path between two neighboring taxa, e.g., 1 and 2. To trace back the optimal 3-path, we define the index matrix (α^3_{uv}):

$$(\alpha^3_{uv}) = \begin{pmatrix} - & - & 2 & 3 & 3 \\ & - & - & 3 & 4 \\ & & - & - & 4 \\ & & & - & - \\ & & & & - \end{pmatrix},$$

where α^3_{uv} denotes the next to last taxon label on the longest ordered 3-path between taxon u and v. Thus the longest ordered 3-path from taxon 1 to taxon 3 contains taxon 2, and the longest ordered 3-path from taxon 2 to taxon 5 contains taxon 4, etc.

Finally we calculate the lengths of all longest circular 3-tours between pairs of taxa i and j by solving equation (4). We get:

$$(L^3_{uv} + d_{uv}) = \begin{pmatrix} - & - & 21 & 28 & 27 \\ & - & - & 25 & 27 \\ & & - & - & 23 \\ & & & - & - \\ & & & & - \end{pmatrix}$$

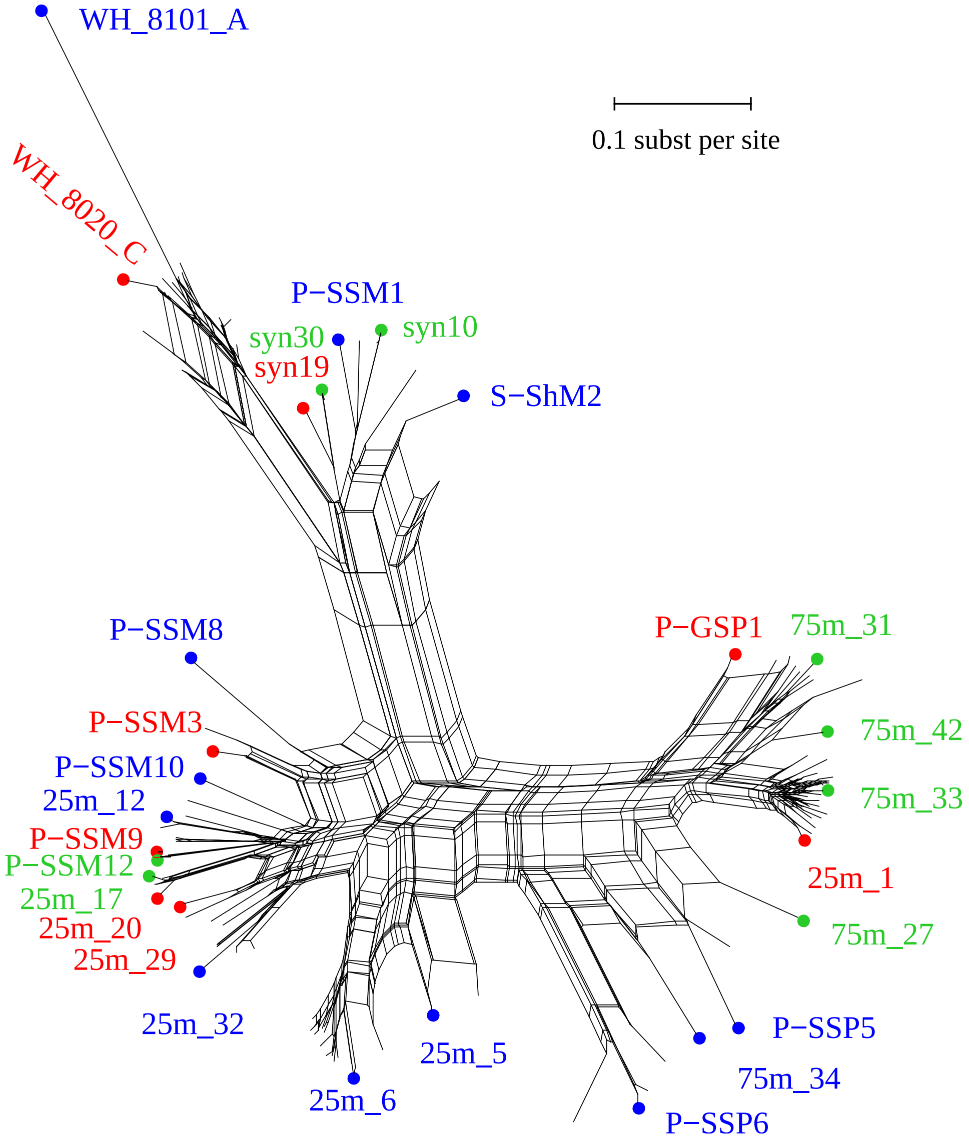


Figure 4: The neighbor-net of the CYANO data with taxa in the union of PD_{20}^{NJ} and PD_{20}^{NNet}. The taxa colors are coded in the same way as in Figure 3.

For the maximal elements of this matrix we can construct the underlying PD_3 sets. Thus each taxon set in PD_3 has a score of $pd_{\max}(3) = 28/2 = 14$ and taxa 1 and 4 span the longest circular 3-tour. To recover the taxon set establishing the longest ordered 3-path from 1 to 4, we simply look at the stored index α_{14}^3 and recognize that it is 3. Therefore, a (and actually the only) PD_3 set is $\{1, 3, 4\}$ with score 14.

Modifications for networks with outgroups

We have introduced `PD-NET` on undirected circular split systems. However, if one can distinguish a circularity-retaining outgroup, i.e. a root taxon can uniquely be placed on the split separating the outgroup from the rest of the taxa, `PD-NET` can easily be extended to this form of directed networks. Simply label the outgroup as taxon 1. Then we order the remaining taxa according to the circular order induced by the underlying network. Now we want to compute $pd_{\max}(k)$ including the outgroup, which is accomplished by solving the problem below:

> *Let n taxa be in a circular order $(1, 2, \ldots, n)$ with pairwise split-distances (d_{uv}) and a number $k < n$. Find the longest circular k-tour originating in the first taxon.*

Conceptually, this is a special case of the problem for unrooted circular split systems with the restriction that every ordered k-path starts at the root taxon. The dynamic programming algorithm remains applicable since the main argument is the same: If $S = (1, s_2, \ldots, s_k, 1)$ is the longest circular k-tour, then $(1, s_2, \ldots, s_i)$, $i = 2, \ldots, k$, is the longest ordered i-path from 1 to s_i. The algorithm proceeds in the same way as for the unrooted case by computing $L^2_{1v}, L^3_{1v}, \ldots, L^k_{1v}$ and then obtains $PD(S)$ based on L^k_{1v}. However, the computational complexity is reduced by the factor n (see appendix).

Analysis of Real Data

To illustrate the difference between a tree based and a network based computation we analyzed two datasets. The first dataset (CYANO) consists of 112 cyanobacteria, cyanophages (cyanobacterial viruses), and environmental taxa (Sullivan et al., 2006). Although the paper does not deal with bacterial diversity, the data are useful to demonstrate the different outcome of *PD*-computation since horizontal gene transfer has occurred among the cyanobacteria, in which the cyanophages play a key role (Sullivan et al., 2006). We retrieved the *psbA* gene data (the core photo system reaction center genes) for all taxa from the NCBI GenBank (Benson et al., 2006). The corresponding sequences were aligned using ClustalW (Thompson et al., 1994) producing an alignment with 729 positions.

The second dataset (CRAYF) comprises the freshwater crayfish of Australia (Shull et al., 2005). The alignment combines four genes: the mitochondrial *16S* rDNA, *12S* rDNA, and *COI* genes and the nuclear *28S* gene. Among 129 aligned sequences, 20 of them had genetic distance of zero to at least one other sequence. They were excluded resulting in an alignment with 109 taxa. The conservation priorities based on the crayfish phylogeny were previously studied in Whiting et al. (2000). However, the conservation priorities were solely based on the *16S* rDNA gene and a subset of 35 crayfish was analysed.

For the CYANO data we calculated the pairwise maximum likelihood distances using IQPNNI (Minh et al., 2005) under the HKY85+G model of substitution (Hasegawa et al., 1985; Yang, 1994). For the CRAYF we used the GTR+I+G model (Tavaré, 1986;

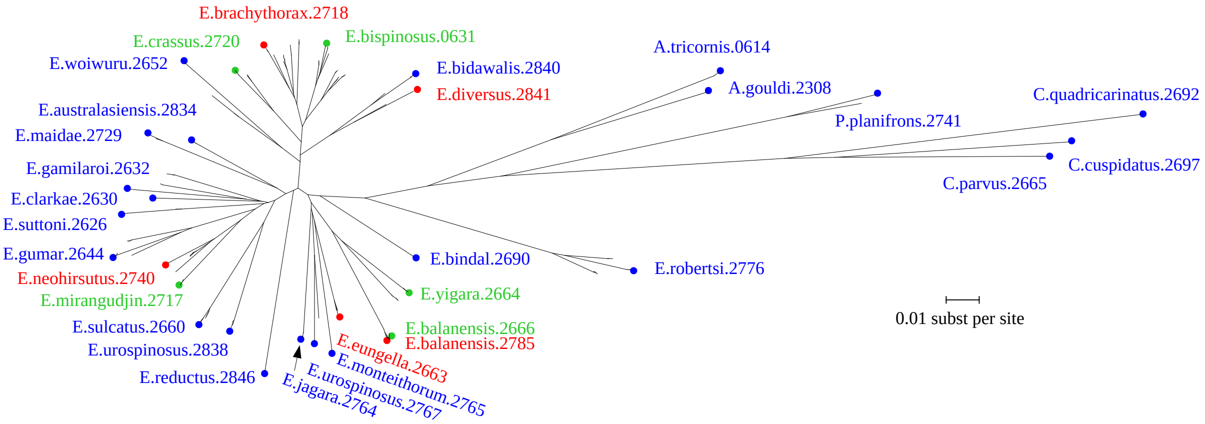


Figure 5: The neighbor-joining tree of the CRAYF data with taxa in the union of PD_{27}^{NJ} and PD_{27}^{NNet}. The taxa colors are coded in the same way as in Figure 3.

Yang, 1994; Gu et al., 1995). The chosen models were suggested from the corresponding papers. The model parameters were estimated from the maximum likelihood tree reconstructed by IQPNNI. For the two resulting distance matrices we computed the neighbor-joining tree (NJ; Saitou and Nei, 1987) and the neighbor-net (NNet; Bryant, 2004) using the program SplitsTree 4 (Huson and Bryant, 2006). Finally, we applied the `PD-NET` to the distances derived from the NJ and NNet. These matrices were used to compute the PD_k^{NJ} and PD_k^{NNet} sets for a given k.

Results

Figure 3 and 4 show the NJ tree and the NNet constructed from the CYANO data, respectively. We only show the results for $k = 20$, thus we want to conserve a bit less than 20% of the taxa. To avoid a crowded illustration only taxa are shown that occur at least in one of the PD-set. The blue taxa appear in both sets, the red taxa are only optimal on the NJ tree while the green taxa are exclusive to the NNet.

First of all, we notice that the structure of the NJ tree and the NNet do, by and large, agree. The circular order of the labels is in both "phylogenies" preserved, except for the taxon $25m\_12$. The corresponding PD sets overlap in 12 taxa (core-taxa) and 8 taxa occur exclusively in one or the other PD-set.Thus the discrepancy of the taxa representing the PD-set is considerable. However, the disagreement is not "evenly" distributed. Subtree B (Figure 3) is not represented by a blue core taxon, i.e. the methods cannot agree on a representation of this subtree. In such a case the conservation decision depends on the reconstruction method. Subtree C (Figure 3) displays nicely the effect of the NNet on displaying genetic relatedness. The PD_{20}^{NJ} set comprises three blue core taxa but taxon $75m\_27$ is not included in the tree based PD-set. However, looking at the corresponding position in the NNet graph it becomes obvious that $75m\_27$ occupies a more intermediate position between group B and group C. Because of this intermediate

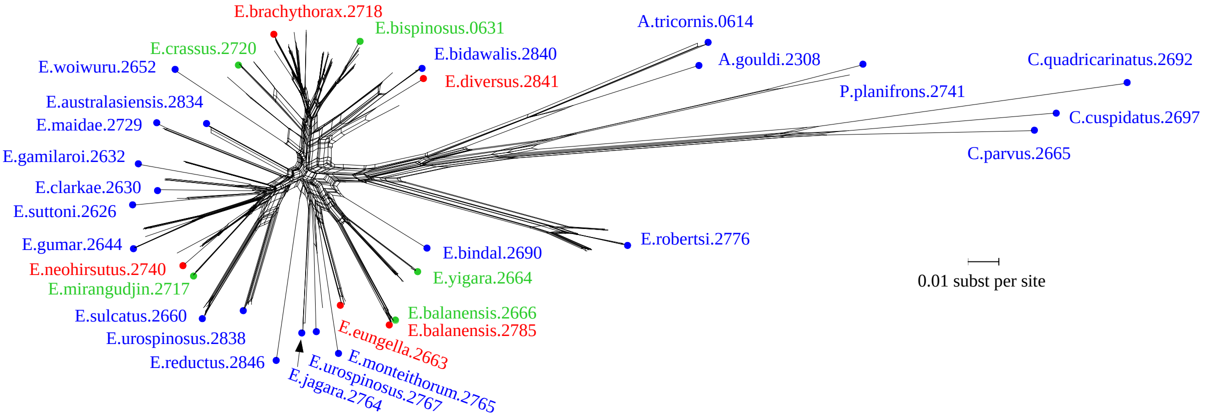


Figure 6: The neighbor-net of the CRAYF data with taxa in the union of PD_{27}^{NJ} and PD_{27}^{NNet}. The taxa colors are coded in the same way as in Figure 3.

position it contributes substantially to the diversity and should be included in the data. Incidentally, such intermediates may be really worthwhile protecting.

For the CRAYF data (Figures 5 and 6) the discrepancy is less striking. For $k = 27$, we observe 22 core taxa and only five taxa that occur exclusively in one of the phylogenies. The circular order of the taxa is perfectly maintained and the tree and the network is not well resolved, i.e. it is difficult to delineate subgroups in the tree. Thus, the CRAYF data are an instance where a NNet analysis does not provide more information than a NJ analysis . However, we would like to emphasize, that this result is not known a priori. The `PD-NET` algorithm may help also in such cases, simply to support the results of a PD-set computation based on a tree only.

Discussion

The selection of single genes can provide the user with ways to guide the conservation analysis into a preferred direction. In this light, conservation studies can be declared to be subjective if only few genetic information is involved in the determination of taxa. Moreover, even the inclusion of several genes to generate a supertree or consensus tree will lead to a loss of diversity information namely conflicting signals.

We have presented an alternative framework to evaluate the diversity of a taxon set and to generate optimal taxon sets of a predefined size k on circular split systems. Such split systems permit the user to incorporate conflicting signals. To this end, we illustrated our method on two real-world datasets, one with horizontal gene transfer and the other with four distinct genes. We compared the two optimal PD sets inferred from the neighbor-joining tree and the neighbor-net. The results indicate that different ways to summarize genetic diversity influence the determination of optimal taxon sets. We expect that the `PD-NET` will be able, if different tree distance matrices are combined and a NNet is computed from the resulting mixture of distance matrices.

Moreover, the discrepancy between optimal *PD* sets from trees and networks gives the user additional information for reserve selection. In particular, the user could fix a size k but optimize for a larger value $\kappa > k$ for both cases and thus can accumulate a huge variety of alternative taxa to be added in later stages. The actual ordering of the reserves could be achieved by employing, e.g., the *Shapley index* (Haake et al., 2005).

Concerning the `PD-NET` algorithm, it is worth noting that in a directed acyclic graph, finding the longest path can be solved by a dynamic programming algorithm (Sedgewick, 2002). Here we extend the problem to the longest k-path and show that the dynamic programming strategy still works. This not only contributes a theoretical result to combinatorial optimization but also shows a direct application of the method to the phylogenetic diversity maximization.

We also measured the performance of the algorithm on a big dataset based on the *rbcL* gene containing 736 flora from a biodiversity hotspot, the Cape of South Africa (Forest et al., 2007). We repeated the same procedure as described in the data analysis section to build the NJ and NNet. On a 2.2GHz computer, the algorithm consumed 25 seconds to compute all optimal PD_k^{NJ} and PD_k^{NNet} sets for $k = 2, \ldots, 736$. The provided algorithm is therefore suitable for most applications that deal with hundreds of taxa or more.

Pardi and Goldman (2005) showed that phylogenetic diversity can also be employed to prioritize species for sequencing in genomics. With the observation that the evolutionary history of bacteria and plants (e.g., Suerbaum et al., 2001; de las Rivas et al., 2004; Hertel et al., 2006) is usually visualized by networks, this presents an additional field to apply our algorithm. In this context, one has to carefully consider whether a tree or a network should be used as the underlying evolutionary model. Applications in comparative genomics usually start with a predefined set of taxa and extend this set with an optimal selection from the remaining taxa. This option is included in our algorithm. We also allow the user to look for alternative taxon sets in PD_k. However, this task depends on the number of PD_k sets and can lead to exponential computing time in the worst case.

Recently, alternative measures for diversity were discussed. Budget constraints as described in the *Noah's Ark Problem* (Weitzman, 1998; Hartmann and Steel, 2006; Pardi and Goldman, 2007) receive an increased interest. Here, an overall budget is prescribed to signify the conservation effort. For each taxon a sub-budget is assigned as the requirement for its survival. The optimization of diversity is now influenced by the sustainability of the optimal set, i.e., we look for a taxon collection whose preservation costs do not exceed the allotted budget. Such a model is clearly not restricted to trees but can also be extended to networks.

Computer Program

A computer program written in C++ is freely available from `http://www.cibiv.at/software/pda/`. The program accepts the input file in NEXUS format (e.g., as produced by SplitsTree) as well as the tree file in NEWICK format. We integrate NEXUS Class Library (NCL; Lewis, 2003) to parse the NEXUS file. The software will then automatically detect the type of the input file to apply appropriate *PDA* algorithms.

Acknowledgements

We would like to thank Heiko Schmidt and Ross Crozier for fruitful discussions. Moreover, we thank Félix Forest and Keith Crandall for providing their data sets. Fabio Pardi pointed out an improvement for the memory complexity of the method. Steffen Klaere and Arndt von Haeseler would like to thank the Isaac Newton Institute for their hospitality while the manuscript developed. Financial support from the Wiener Wissenschafts-, Forschungs- and Technologiefonds (WWTF) is greatly appreciated.

Appendix: Ordered k-paths

We show, that if $(s_1, s_2, \ldots, s_k, s_1)$ is the longest circular k-tour then $(s_1, s_2, \ldots, s_k)$ is the longest ordered k-path from s_1 to s_k. This is easily proven by assuming that there is a longer ordered k-path $(s_1, t_2, \ldots, t_{k-1}, s_k)$ from s_1 to s_k. Then

$$L(s_1, t_2, \ldots, t_{k-1}, s_k) + d_{s_1 s_k} > L(s_1, s_2, \ldots, s_k) + d_{s_1 s_k},$$

and therefore $(s_1, s_2, \ldots, s_k, s_1)$ would no longer be the longest circular k-tour. Similarly, if $(s_1, s_2, \ldots, s_k)$ is the longest ordered k-path from s_1 to s_k then also $(s_1, s_2, \ldots, s_{k-1})$ is the longest ordered $(k-1)$-path from s_1 to s_{k-1}. We say, that our problem exhibits an *optimal substructure* (Cormen et al., 2001), i.e., the longest circular k-tour $(s_1, s_2, \ldots, s_k, s_1)$ contains also the longest ordered i-path from s_1 to s_i for all $i = 2, \ldots, k$. Therefore, we can apply a *dynamic programming algorithm* to construct all the longest ordered i-paths between every pair of taxa and subsequently combine them to accomplish a longest circular k-tour.

Appendix: Complexity considerations

Listings 1 and 2 describe the algorithm for the unrooted case in pseudo-code. Listing 1 calculates $k-1$ matrices (L^i_{uv}) and (α^i_{uv}) for i from 2 to k given the circular order and the split-distance matrix (d_{uv}). Listing 2 constructs an optimal k-set based on the computed information (L^i_{uv}) and (α^i_{uv}) by identifying two taxa $\tilde{u}$ and $\tilde{v}$ maximizing (4) and incrementally adding the taxon maximizing (5) into the set.

The computation of matrices (L^i_{uv}) and (α^i_{uv}) needs four nested loops as shown in Listing 1. The three outer loops generate all possible combinations of (i, u, v) which amounts to $O(kn^2)$ since i runs from 3 to k, while u and v run from 1 to n. The fourth loop has complexity $O(n)$ because the taxon index s varies between u and v. We get the cumulative time complexity of $O(kn^3)$. The computation of the optimal k-set as seen from listing 2 requires $O(n^2)$ time for the determination of two taxa $\tilde{u}$ and $\tilde{v}$ and $O(k)$ time for identifying $k-2$ remaining taxa. In total, the computational complexity of the algorithm is $O(kn^3)$ in the unrooted case. In the rooted case the time complexity reduces to $O(kn^2)$.

Considering memory requirement, one observes the following property of equation (5). Each row of the matrix (L^i_{uv}) is computed using only the same row of (L^{i-1}_{uv}) and the split-distance matrix (d_{uv}). Hence, one can compute the first rows (L^i_{1v}) and (α^i_{1v}) and infer the longest circular k-tour originating at taxon 1. Subsequently, one can re-use the memory space to calculate the longest k-tour starting at taxon u, $u = 2, \ldots, n-k+1$. With this trick, the memory requirement for the unrooted case can be reduced to $O(kn)$ which is also the memory complexity of the rooted case.

Listing 1: Compute L^i_{uv} and α^i_{uv}

```
Input: Set of taxa $X = \{1, 2, \ldots n\}$, indexed in a circular order;
Subset size $k$;
Split-distance matrix $(d_{uv})$
Output: Lengths $L^i_{uv}$ of longest ordered $i$-paths between all pairs of taxa $u, v$, for
        $1 \leq u < v \leq n$, $2 \leq i \leq k$;
Indices $\alpha^i_{uv}$ of the taxon set which generates $L^i_{uv}$
begin
    Init $L^2_{uv} = d_{uv}$ for all $1 \leq u < v \leq n$;
    for $i = 3$ to $k$ do
        for $u = 1$ to $n - i + 1$ do
            for $v = u + i - 1$ to $n$ do
                Init $L^i_{uv} = 0$;
                for $s = u + i - 2$ to $v - 1$ do
                    if $L^i_{uv} < L^{i-1}_{us} + d_{sv}$ then
                        Update $L^i_{uv} = L^{i-1}_{us} + d_{sv}$;
                        $\alpha^i_{uv} = s$;
end
```

Listing 2: Construct an optimal k-set

```
Input: Set of taxa $X = \{1, 2, \ldots n\}$;
Subset size $k$;
Split-distance matrix $(d_{uv})$;
Matrices $(L^i_{uv})$ and $(\alpha^i_{uv})$
Output: Set $S$ of $k$ taxa with maximal $PD$
begin
    max = 0;
    for $u = 1$ to $n - k + 1$ do
        for $v = u + k - 1$ to $n$ do
            if max $< L^k_{uv} + d_{uv}$ then
                Update max $= L^k_{uv} + d_{uv}$;
                $\tilde{u} = u$;
                $\tilde{v} = v$;
    Init $S = \{\tilde{u}, \tilde{v}\}$;
    for $i = k$ downto 3 do
        $S = S \cup \{\alpha^i_{\tilde{u}\tilde{v}}\}$;
        Set $\tilde{v} = \alpha^i_{\tilde{u}\tilde{v}}$;
    return $S$;
end
```

References

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

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2006. GenBank. Nucl. Acids Res. 34:D16–D20.

Bryant, D. 2004. The splits in the neighborhood of a tree. Ann. Combinatorics 8.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2001. Introduction to Algorithms. 2nd ed. MIT Press and McGraw-Hill.

Crozier, R. H. 1992. Genetic diversity and the agony of choice. Biol. Conserv. 61:11–15.

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

de las Rivas, B., Á. Marcobal, and R. Muñoz. 2004. Allelic diversity and population structure in *oenococcus oeni* as determined from sequence analysis of housekeeping genes. Applied and Environmental Microbiology 70:7210–7219.

Degnan, J. H. and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Doolittle, W. F., Y. Boucher, C. L. Nesbo, C. J. Douady, J. O. Andersson, and A. J. Roger. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos Trans R Soc Lond B Biol Sci. 358:39–57.

Dress, A. and D. Huson. 2004. Constructing splits graphs. IEEE/ACM Trans. Comput. Biol. Bioinform. 1:109–115.

Dress, A. W. M., K. T. Huber, and V. Moulton. 2001. Totally split-decomposable metrics of combinatorial dimension two. Ann. Combin. 5:99–112.

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. Evolutionary Biology Online 2:70–77.

Forest, F., R. Grenyer, M. Rouget, T. J. Davies, R. M. Cowling, D. P. Faith, A. Balmford, J. C. Manning, S. Proches, M. van der Bank, G. Reeves, T. A. J. Hedderson, and V. Savolainen. 2007. Preserving the evolutionary potential of floras in biodiversity hotspots. Nature 445:757–760.

Gaston, K. J. and J. I. Spicer. 2004. Biodiversity: An Introduction. 2nd ed. Blackwell Publishing Professional.

Graur, D. and W.-H. Li. 2000. Fundamentals of Molecular Evolution. 2nd ed. Sinauer Associates, Sunderland, Massachusetts.

Gu, X., Y.-X. Fu, and W.-H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12:546–557.

Haake, C.-J., A. Kashiwada, and F. E. Su. 2005. The Shapley value of phylogenetic trees. Working Paper 363 Institute of Mathematical Economics, Bielefeld.

Hartmann, K. and M. Steel. 2006. Maximizing phylogenetic diversity in biodiversity conservation: Greedy solutions to the noah's ark problem. Syst. Biol. 55:644–651.

Hasegawa, M., H. Kishino, and T.-A. Yano. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Henz, S. R., D. H. Huson, A. F. Auch, K. Nieselt-Struwe, and S. C. Schuster. 2005. Whole-genome prokaryotic phylogeny. Bioinformatics 21:2329–2335.

Hertel, J., M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I. L. Hofacker, P. F. Stadler, and T. S. of Bioinformatics Computer Labs 2004/2005. 2006. The expansion of the metazoan microRNA repertoire. BMC Genomics 7.

Huber, K. T. and V. Moulton. 2005. Phylogenetic networks. Pages 178–204 *in* Mathematics of Evolution and Phylogeny (O. Gascuel, ed.). Oxford University Press, Oxford, UK.

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

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

Lewis, P. O. 2003. NCL: a C++ class library for interpreting data files in NEXUS format. Bioinformatics 19:2330–2331.

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

Minh, B. Q., L. S. Vinh, A. von Haeseler, and H. A. Schmidt. 2005. pIQPNNI – parallel reconstruction of large maximum likelihood phylogenies. Bioinformatics 21:3794–3796.

Moulton, V., C. Semple, and M. Steel. 2007. Optimizing phylogenetic diversity under constraints. Journal of Theoretical Biology 246:186–194.

Nee, S. and R. M. May. 1997. Extinction and the Loss of Evolutionary History. Science 278:692–694.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

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

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

Saitou, N. and M. Nei. 1987. The neighbor–joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Sedgewick, R. 2002. Algorithms in C++ Part 5: Graph Algorithms. Addison Wesley.

Shull, H. C., M. Pérez-Losada, D. Blair, K. Sewell, E. A. Sinclair, S. Lawler, M. Ponniah, and K. A. Crandall. 2005. Phylogeny and biogeography of the freshwater crayfish *Euastacus* (Decapoda: Parastacidae) based on nuclear and mitochondrial DNA. Molecular Phylogenetics and Evolution 37:249–263.

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

Suerbaum, S., M. Lohrengel, A. Sonnevend, F. Ruberg, and M. Kist. 2001. Allelic diversity and recombination in campylobacter jejuni. Journal of Bacteriology 183:2553–2559.

Sullivan, M. B., D. Lindell, J. A. Lee, L. R. Thompson, J. P. Bielawski, and S. W. Chisholm. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLOS Biology 4:1344–1357.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences 17:57–86.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions–specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

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. 1998. The Noah's Ark problem. Econometrica 66:1279–1298.

Whiting, A. S., S. H. Lawler, P. Horwitz, and K. A. Crandall. 2000. Biogeographic regionalization of Australia: assigning conservation priorities based on endemic freshwater crayfish phylogenetics. Animal Conservation 3:155–163.

Wilson, E. O., ed. 1997. Biodiversity. 2nd ed. National Academies Press.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximative methods. J. Mol. Evol. 39:306–314.