Il Buono, Il Brutto, Il Cattivo

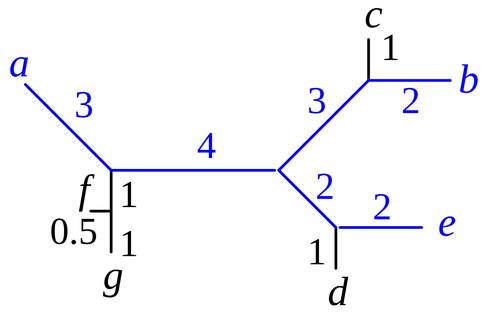
Nguyen Thi Minh Anh, Bui Quang Minh, Arndt von Haeseler and Steffen Klaere

Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories,
University of Vienna, Medical University of Vienna, Veterinary University of Vienna

Introduction

Phylogenetic Diversity

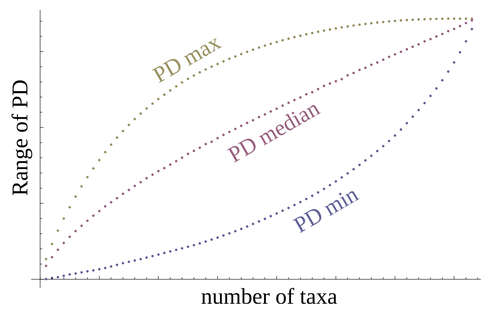
- Simplest approach to evaluate the diversity of a habitat is to count the species observed in the area. Associated measure is called *taxon richness*.
- Faith [1992] introduced a measure which incorporates evolutionary relationships between taxa. To this end, the genetic information available for a given set of taxa $\mathcal{X}$ is used to reconstruct a phylogenetic tree $\mathcal{T}$. The *phylogenetic diversity* of a subset $Y \subset \mathcal{X}$ is given as the sum of the length of the branches connecting Y in $\mathcal{T}$.



For the above tree, compare the taxon sets $S_1 = \{a, b, e\}$ and $S_2 = \{b, e, f, g\}$. In terms of taxon richness, set S_2 is prioritized while phylogenetic diversity prioritizes S_1 due to a score of 16 compared to 15.5 for S_2.

The Range of PD

The chart below indicates the advantage of PD over taxon richness. For a given number k of taxa, taxon richness treats all sets of this size equally while PD assigns values in the full range between the best and the worst score possible.



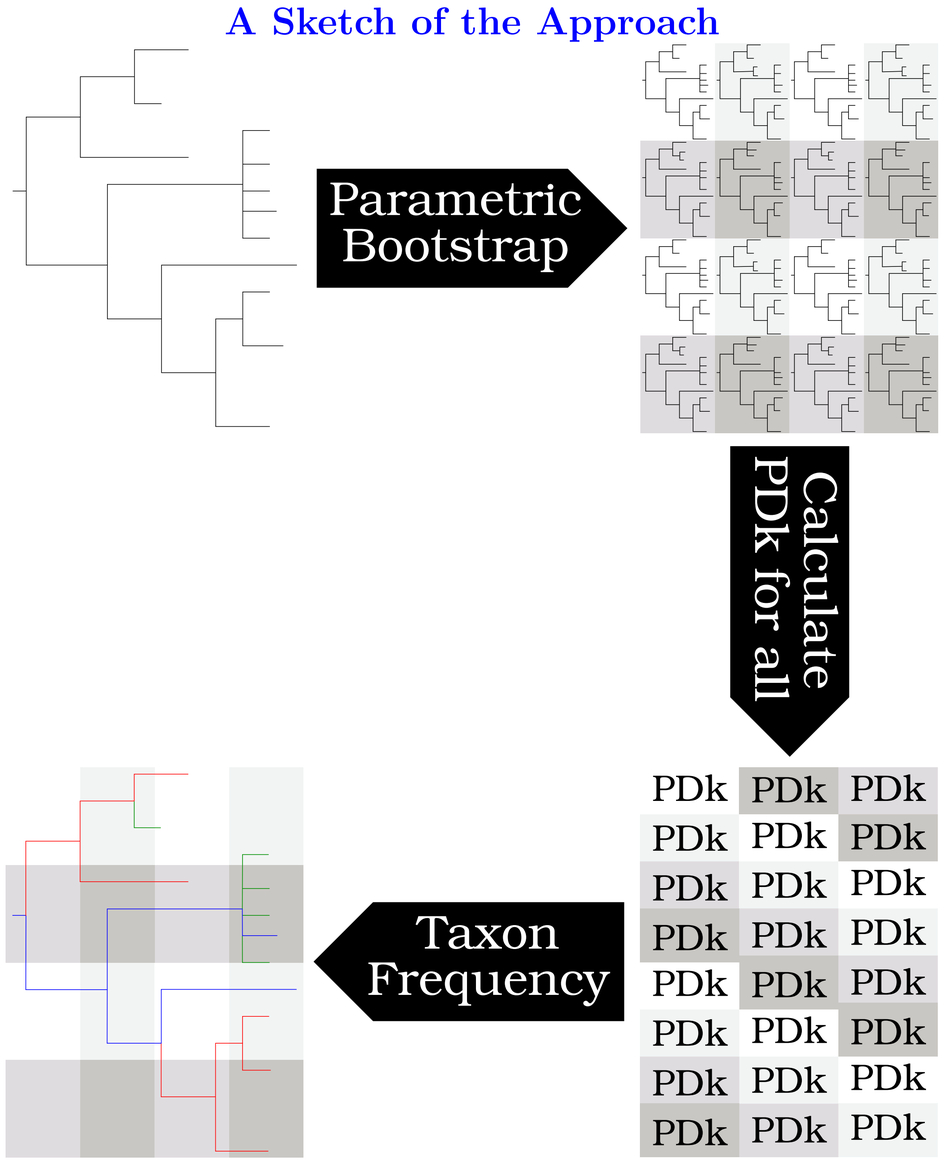
Optimization Strategies

Steel [2005], Pardi and Goldman [2005] posed the following optimization problem:
Given a set of taxa $\mathcal{X}$ and a phylogeny $\mathcal{T}$ connecting them. Find for any number k the subset $\mathrm{PD}_k^{\max}$ maximizing the PD score for all sets of size k. This distinction is the basis of the following approach.

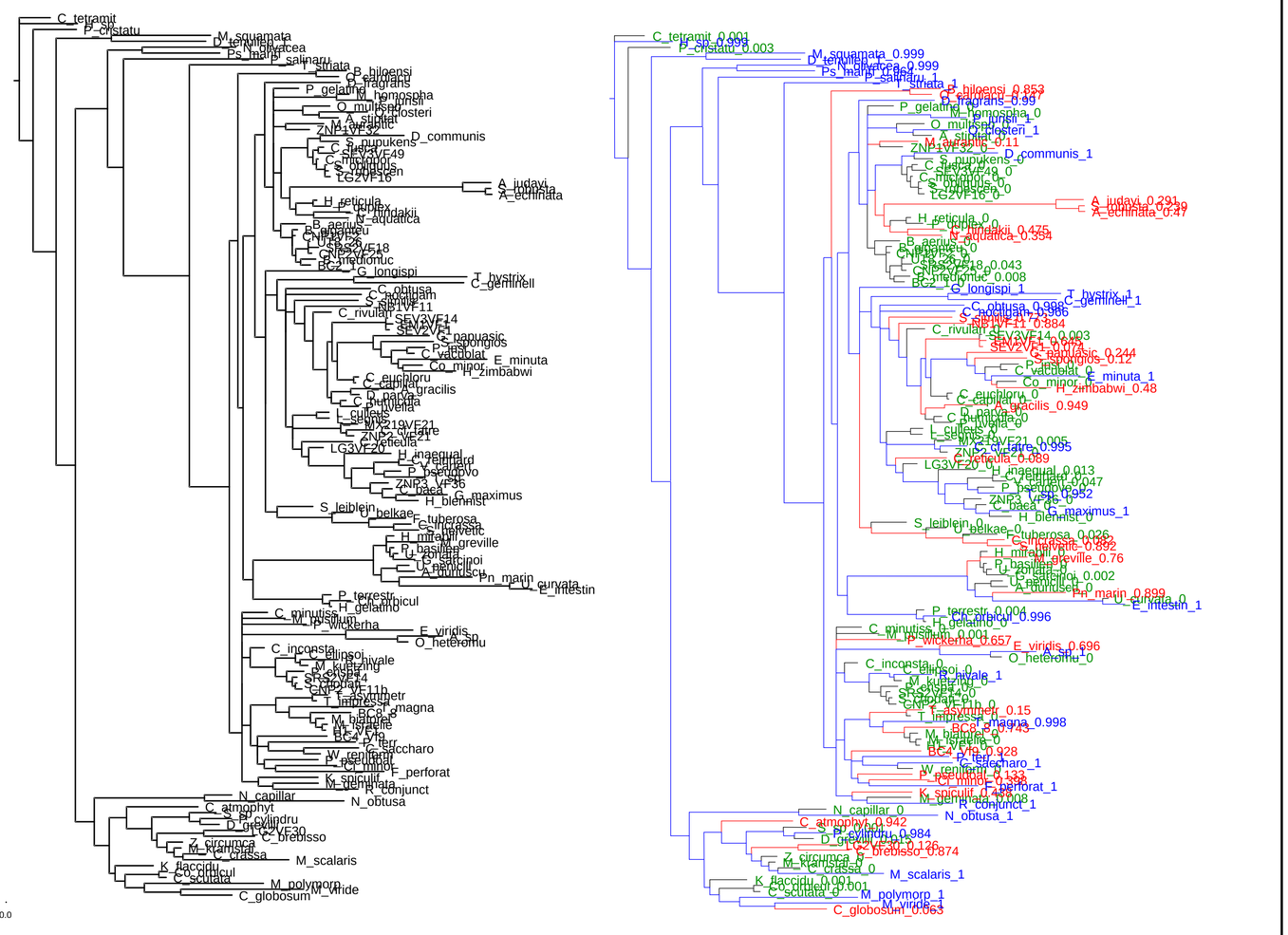
Some Challenges

- What is the actual distribution of PD?
  - Forest et al. [2007] estimated the distribution of PD for every number of taxa k by generating 25.000 random samples of k taxa and calculating their PD. They used this information to show that PD is more than just a surrogate for taxon richness as most studies so far indicated.
- How do we account for branch length estimation errors?
  - The good, the bad, the ugly approach presented here uses a parametric bootstrap approach to generate a sample of trees. On all trees the optimal PD value for fixed k is calculated. Then, for each taxon the number of occurrences in these sets is counted and it is classified according to the score.
- How do we include conflicting signals in the genetic data?
  - Because the good, the bad, the ugly basically classifies the taxa without looking at the actual PD score, this approach can also be used to look for interesting taxa.
  - On can also consider networks to incorporate conflicts (see other poster).

A Sketch of the Approach



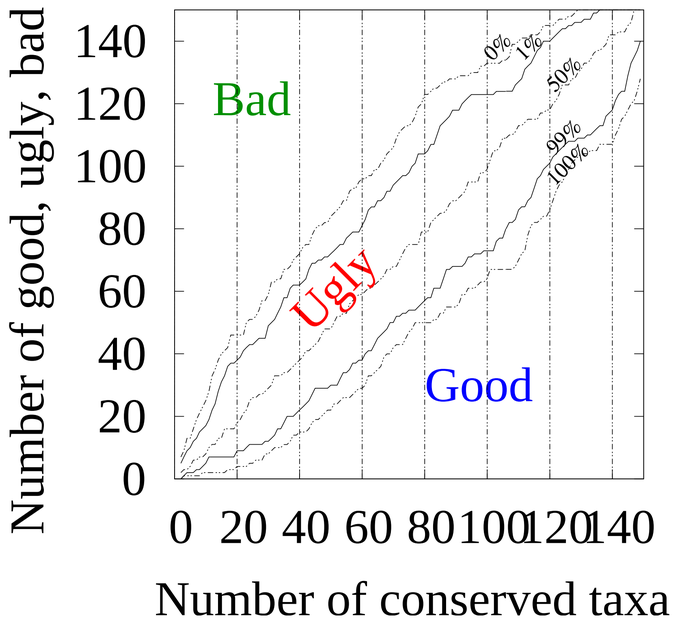
An Example [Lewis and Lewis, 2005]



A tree of 150 taxa of green algae from various ecosystems is reconstructed from 18S rDNA sequences [Lewis and Lewis, 2005]. From the tree presented in the paper we simulated 1000 alignments using SEQ GEN [Rambaut and Grassly, 1997] and generated a tree for each alignment using IQPNNI [Vinh and von Haeseler, 2004]. For all trees the optimal $\mathrm{PD}_{50}^{\max}$ set was obtained using the PDA package [Minh et al., 2006]. If a taxon occured in at least 95% of all PD sets, then it is a **good** taxon, if it occurs in at most 5%, then it is a **bad** taxon, and otherwise it is an **ugly** taxon.

Discussion and Outlook

- Approach indicates the stability of the allegedly optimal choice of taxa by pointing out instable (ugly) choices.
- Approach accounts for errors in branch length estimation. Interesting clades with more than one possible representative taxon are highlighted by a set of ugly taxa.
- Starting with a fixed number of taxa we are given a larger set of taxa to be considered, a valuable approach for taxon reserve selection.
- Closer inspection of ugly taxa to identify clades and select those taxa which are really interesting. One should not select any choice of ugly taxa to complement the set of good taxa. One would rather need a kind of classification of ugly taxa such that selecting a representative from each class completes the set. Often such a class will coincide with a clade. However, there are taxa (like M_aurantic) for which the associated class does not give such an immediate interpretation.
- If one ignores the actual PD score, this approach can also be used to incorporate several trees into the analysis!
- Problem: If a clade has more than 20 potentially ugly taxa then none will get over the 5% threshold.



References

Daniel P. Faith. Conservation Evaluation and Phylogenetic Diversity. *Biol. Conserv.*, 61:1–10, 1992.

Félix Forest, Richard Grenyer, Mathieu Rouget, T. Jonathan Davies, Richard M. Cowling, Daniel P. Faith, Andrew Balmford, John C. Manning, Serban Proches, Michelle van der Bank, Gail Reeves, Terry A. J. Hedderson, and Vincent Savolainen. Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature*, 445:757–760, 2007.

Louise A. Lewis and Paul O. Lewis. Unearthing the molecular diversity of desert soil green algae. *Syst. Biol.*, 54(6):936–947, 2005. ISSN 1063-5157.

Bui Quang Minh, Steffen Klaere, and Arndt von Haeseler. Phylogenetic diversity within seconds. *Syst. Biol.*, 55(5):769–773, 2006. doi: 10.1080/10635150600981604.

Fabio Pardi and Nick Goldman. Species choice for comparative genomics: Being greedy works. *PLoS Genet.*, 1:672–675, 2005. doi: 10.1371/journal.pgen.0010071.

Andrew Rambaut and Nicholas C. Grassly. Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13:235–238, 1997.

Mike Steel. Phylogenetic Diversity and the Greedy Algorithm. *Syst. Biol.*, 54(4): 527–529, 2005.

Le Sy Vinh and Arndt von Haeseler. IQPNNI: Moving fast through tree space and stopping in time. *Mol. Biol. Evol.*, 21:1565–1571, 2004. doi: 10.1093/molbev/msh176.