Tracing phylogenetic signal in datasets

Heiko A. Schmidt

Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories (MFPL)
Vienna, Austria
heiko.schmidt@univie.ac.at

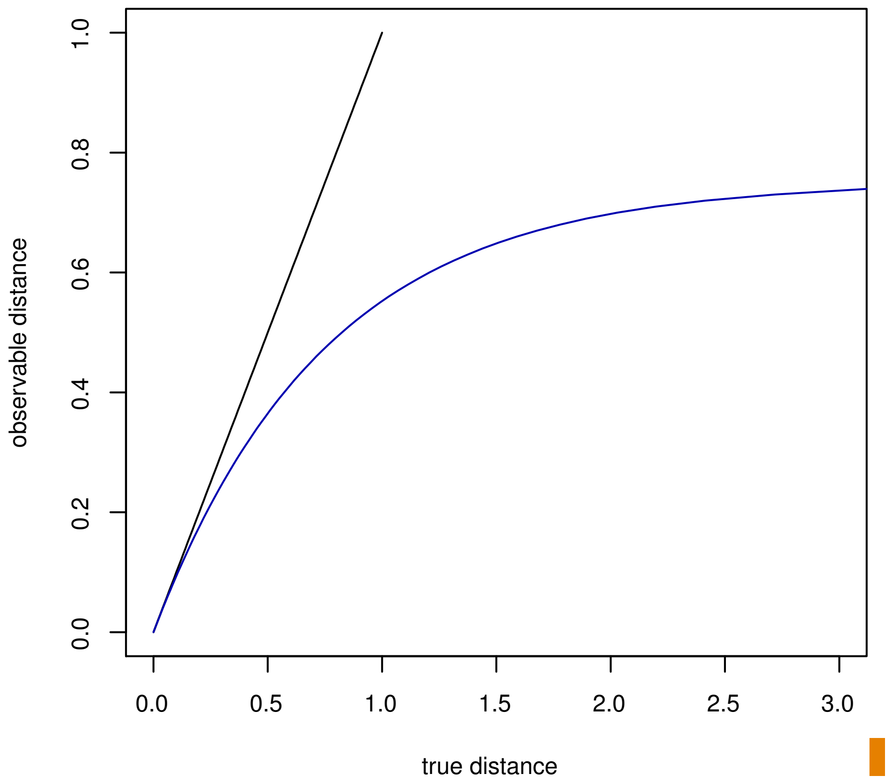
Phylogenetic Information

The information about the true tree, might be obscured or unextractable from an alignment due to

- too similar sequences (no differences → no information)
- sequences are to divergent (saturated sequences → information drowned in noise)

Is there a way to check for this?

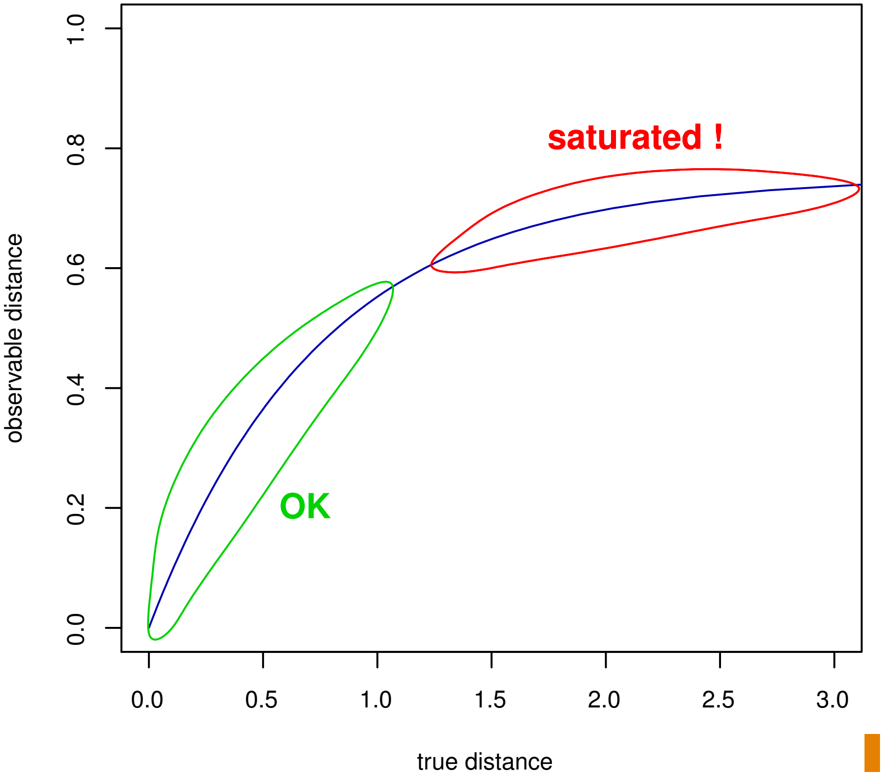
Remember Correction for Multiple Hits



Plotting Mutation Values

- Take every pair of sequences
- Count the number of observable differences (e.g., transitions, transversions)
- Compute the distances of the sequence pair:.
- and plot the distance (x-axis) against the observable differences (y-axis)

Plotting Mutation Values (2)



Posterior Probabilities and Empirical Bayes

- We can now reconstruct ML trees, but how comparable are the likelihoods, how reliable the groupings?

- Branch reliability can be checked, support values computed using:
  - Bootstrapping, Jackknifing alignment columns + consensus.
  - Randomizing input orders in stepwise insertions (TREE-PUZZLE).

Posterior Probabilities and Empirical Bayes

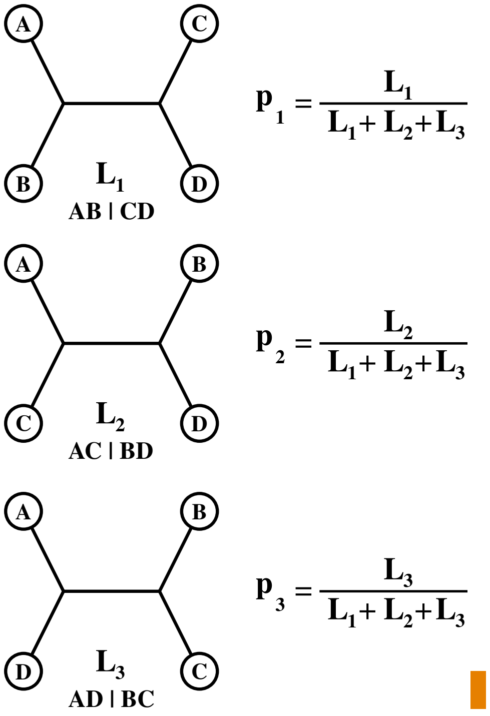
- We have learnt to reconstruct ML trees and heard that one can compare their posterior probabilities. . .

- Problem: How different are likelihoods? Just from the value of likelihoods one often cannot tell whether they are significantly different.

- Nomalization: Posterior probabilities are computed:

$$p_i = \frac{L_1}{\sum_n L_n}$$

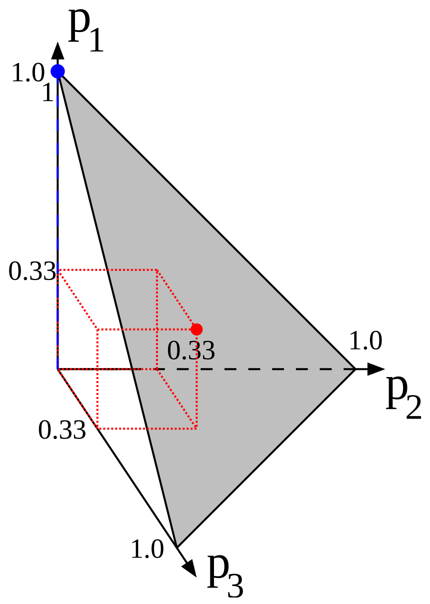
Plotting Posteriors: Likelihood Mapping

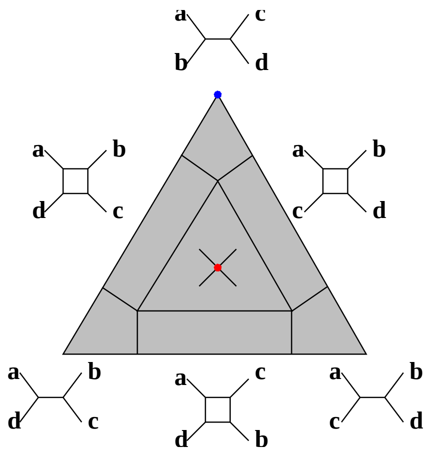


$$p_1 = \frac{L_1}{L_1+L_2+L_3}$$

$$p_2 = \frac{L_2}{L_1+L_2+L_3}$$

$$p_3 = \frac{L_3}{L_1+L_2+L_3}$$

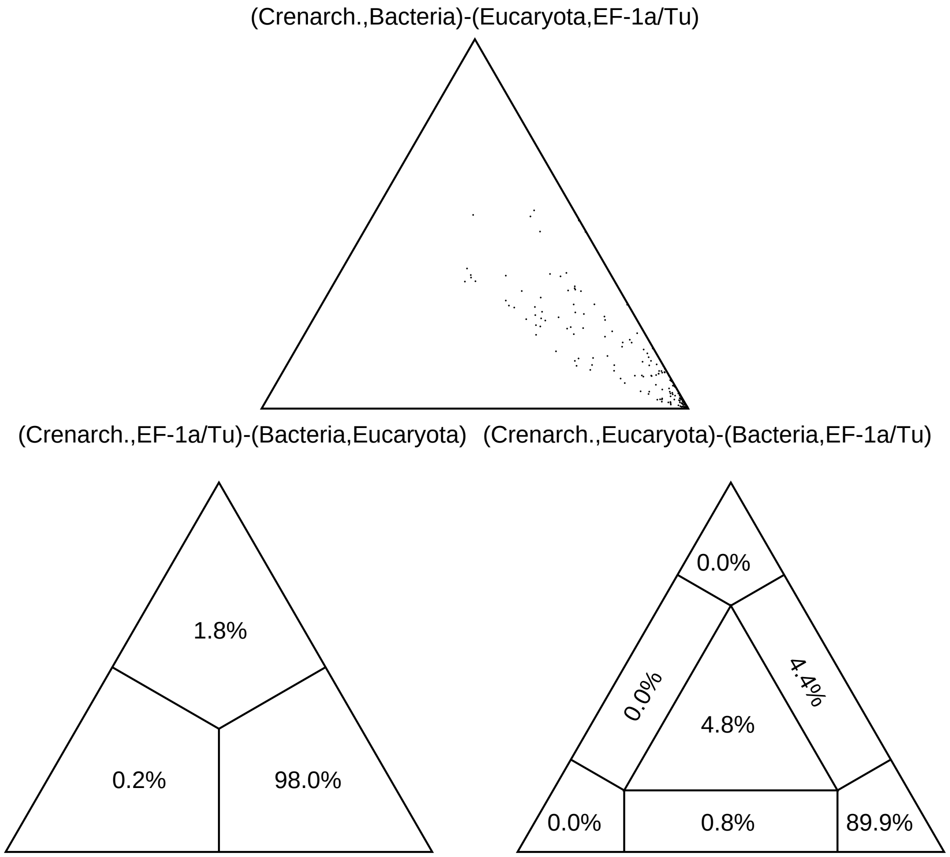




Since $p_1+p_2+p_3 = 1$, 3D points (p_1, p_2, p_3) fall into a triangular (simplex).

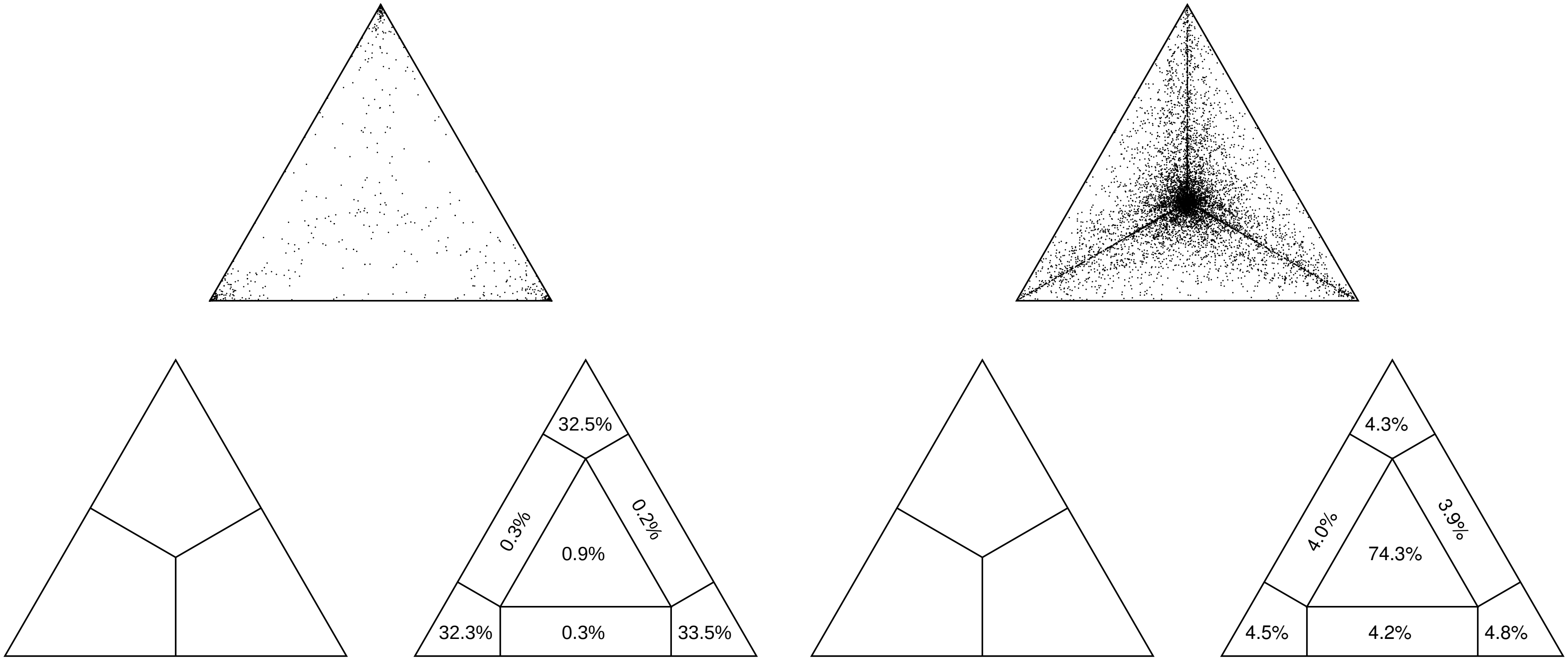
If we repeat this for all quartets (or a large random subset) in a dataset we can assess the amount of phylogenetic signal in the dataset.

Likelihood Mapping (Cluster Analysis)



The Simplex Plot can visualize the relationship among clusters.

Likelihood Mapping (Information Content)



The Simplex Plot can also visualize the information content in an alignment.