

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

September 2008

Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Phylogenetic Information

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

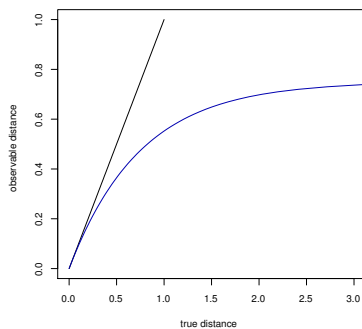
- too similar sequences
(no differences → no information)
- sequences are too divergent
(saturated sequences → information drowned in noise)

Are there ways to check for this?

Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

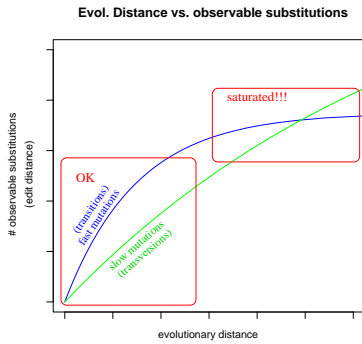
Reminder: Jukes-Cantor Correction for Multiple Hits



Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Plotting Substitutions vs. Distance for DNA



Transitions (ts) usually occur much more often than transversions (tv). Thus, the ts-curve rises faster but reaches the plateau earlier. The tv-curve can only 'overtake' the ts-curve if the latter is quite saturated!

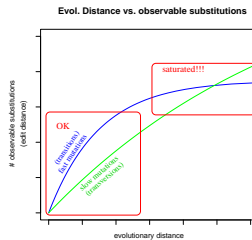
Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Saturation Plots for DNA

Saturation Plots can be created as follows

- Take every pair of sequences
 - Count the number of observable substitutions (e.g., transitions, transversions)
 - Compute the distances of the sequence pair (e.g., with ML)
- ... and plot the evolutionary distance (x-axis) against the observed substitutions (y-axis) for each class of mutations.



Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Software for DNA Saturation Plots

Saturation Plots can be created using

- **Windows:** DAMBE (Xia and Xie, 2001)

→ Graphics menu
→ Transition and transversion vs. divergence

- **All OS:** TREE-PUZZLE (Schmidt *et al.*, 2001), plotting the data in *.tstv with a few lines in the R program (www.r-project.org):

```
tstvtab = read.table("ali.phy.tstv", header=T) # read data
attach(tstvtab) # use headers as names
pdf(file="tstv.pdf") # open PDF file
maxsubst=max(ts,tv) # find maximum
plot(distance,ts,col=2,ylab="observed substitutions",ylim=c(0,maxsubst))
points(distance,tv,col=3) # plot
dev.off() # close PDF file
detach(tstvtab) # release names
q() # quit R program
```

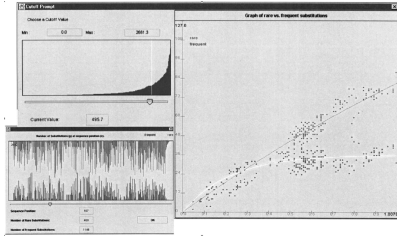
Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Saturation Plots for AA (AsaturA, van de Peer et al. 2002)

The same can be done for amino acids, but

- There is not intuitive division into fast and slow substitutions,
- AsaturA orders the the substitution types according to the probabilities in a substitution probability matrix (e.g., PAM, WAG).
- Then, the user has to set a cutoff between *fast* and *slow*. (But there are no guidelines for that choice.)
- Then the numbers of fast and slow substitutions are plotted against the distance accordingly.



Heiko A. Schmidt Tracing phylogenetic signal in datasets

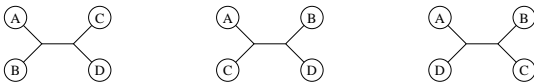
Notes

Likelihood Weights, Posterior Prob., and Empirical Bayes

- We can compute a likelihood value for a tree based given an alignment and model... (cf. the lecture on ML methods).
- Problem: How different are the likelihoods? Just from the value of likelihoods one often cannot tell whether they are significantly different.
- Normalization: Posterior probabilities are computed:

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

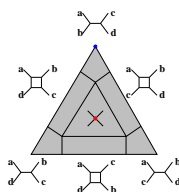
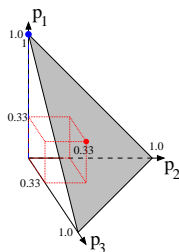
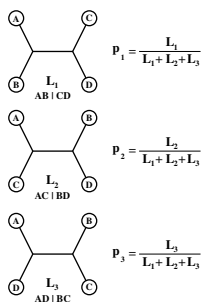
- We can use that on the three different quartet topologies to assess phylogenetic information in our data.



Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Plotting Posteriors: Likelihood Mapping



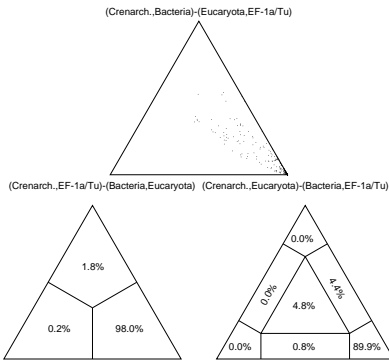
Since $p_1 + p_2 + p_3 = 1$, 3D points (p_1, p_2, p_3) fall into a triangular (simplex).

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.

Notes

Heiko A. Schmidt Tracing phylogenetic signal in datasets

Likelihood Mapping (Cluster Analysis)



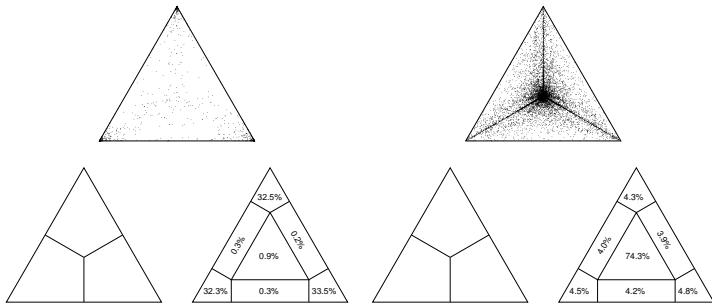
The Simplex Plot can visualize the relationship among (4) sets of taxa.

The taxa/sequences are assigned to four sets (A,B,C,D) one for each leaf of a quartet tree.

Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Likelihood Mapping (Information Content)



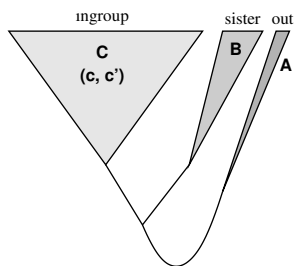
The Simplex Plot can also visualize the information content in an alignment.

By not assigning taxa to clusters, four are chosen randomly for each leaf. We have to add the percentages in the **corners (resolved)** or **rectangles (partly resolved)**, respectively. **Center** means **unresolved**.

Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Likelihood Mapping to Validate Outgroups



- We can check the reliability of an outgroup by assigning taxa to three sets: **C** - the examined ingroup, **B** - an early sister group, and **A** - the outgroup.
- random quartets are drawn from the sets: **two from C** and one each from **B** and **A**.
- if not **a, b|c, c'** (upper corner) is the support topology, **A** is not a good outgroup (or **B** is not a proper sister group).

Heiko A. Schmidt Tracing phylogenetic signal in datasets

Notes

Exercises:

the exercises can be found at

<http://www.cibiv.at/~hschmidt/VEME/signal>

Notes

Notes

Notes
