

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

June 5, 2007

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

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

- too similar sequences
(no differences → no information)

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

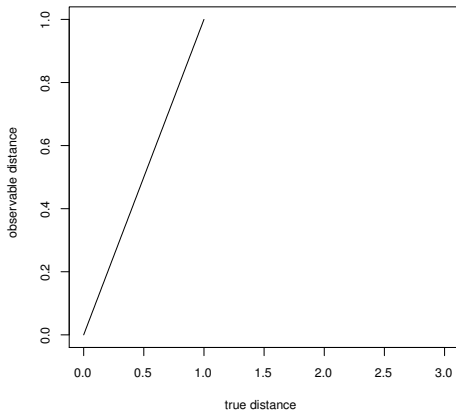
- too similar sequences
(no differences → no information)
- sequences are too divergent
(saturated sequences → information drowned in noise)

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

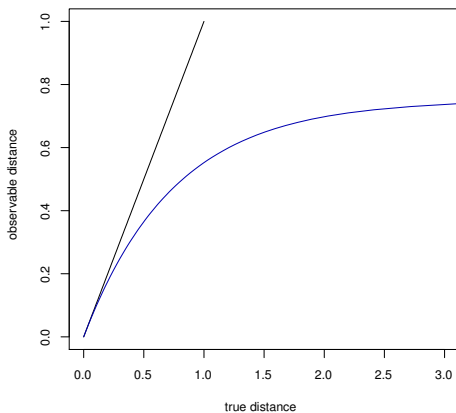
- too similar sequences
(no differences → no information)
- sequences are too divergent
(saturated sequences → information drowned in noise)

Are there ways to check for this?

Reminder: Jukes-Cantor Correction for Multiple Hits

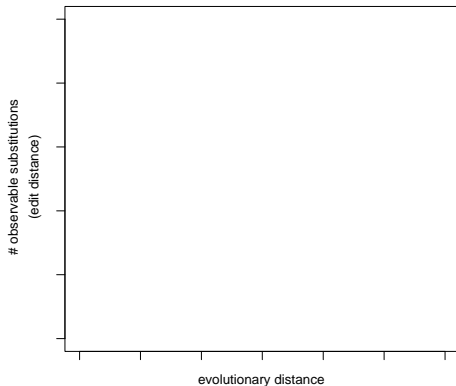


Reminder: Jukes-Cantor Correction for Multiple Hits



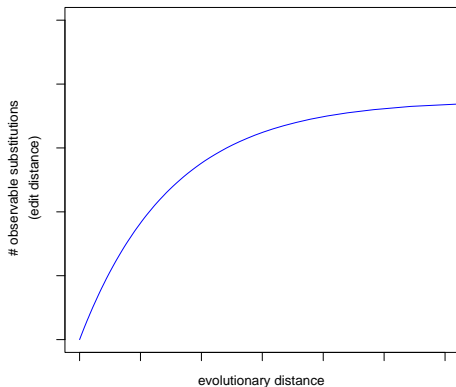
Plotting Substitutions vs. Distance for DNA

Evol. Distance vs. observable substitutions



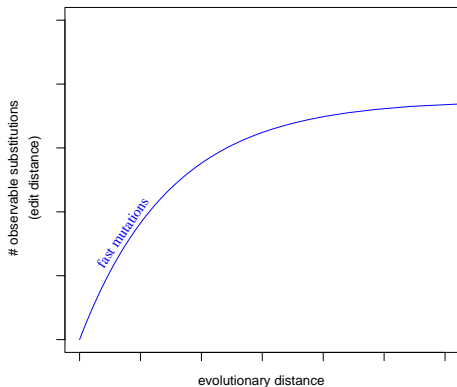
Plotting Substitutions vs. Distance for DNA

Evol. Distance vs. observable substitutions



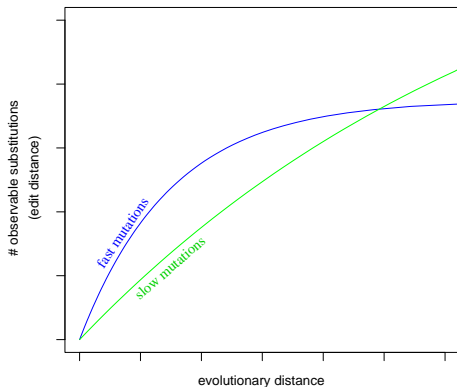
Plotting Substitutions vs. Distance for DNA

Evol. Distance vs. observable substitutions



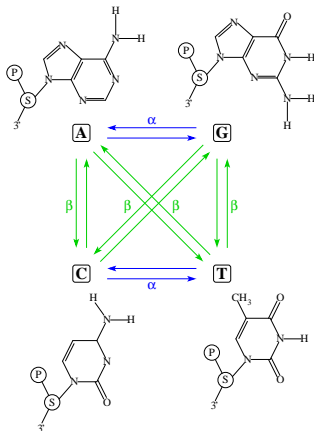
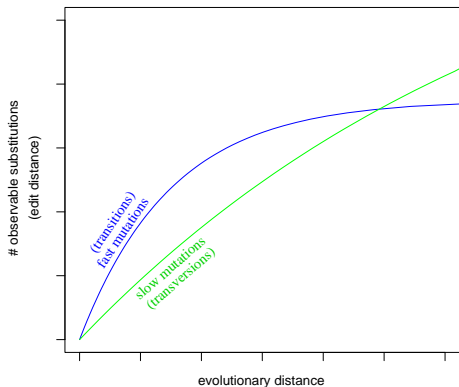
Plotting Substitutions vs. Distance for DNA

Evol. Distance vs. observable substitutions



Plotting Substitutions vs. Distance for DNA

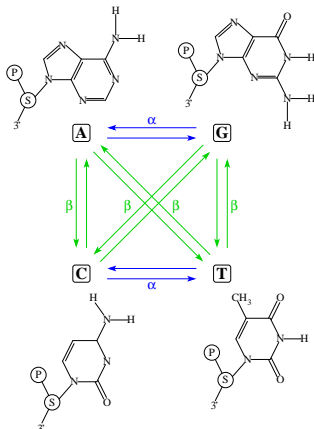
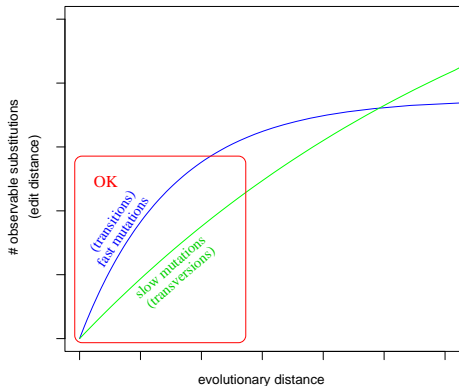
Evol. Distance vs. observable substitutions



Transitions (ts) usually occur much more often than transversions (tv). Thus, the ts-curve rises faster but reaches the plateau earlier.

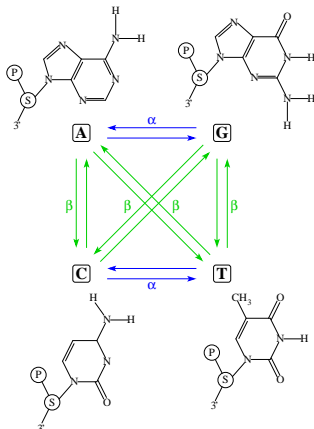
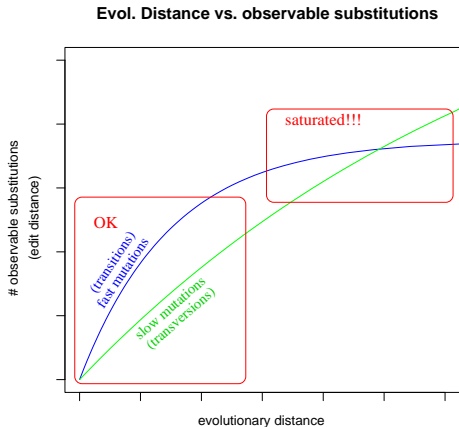
Plotting Substitutions vs. Distance for DNA

Evol. Distance vs. observable substitutions



Transitions (ts) usually occur much more often than transversions (tv). Thus, the ts-curve rises faster but reaches the plateau earlier.

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!

Saturation Plots for DNA in DAMBE

Saturation Plots can be created, e.g., with the DAMBE program as follows

- Take every pair of sequences

Saturation Plots for DNA in DAMBE

Saturation Plots can be created, e.g., with the DAMBE program as follows

- Take every pair of sequences
 - Count the number of observable substitutions (e.g., transitions, transversions)

Saturation Plots for DNA in DAMBE

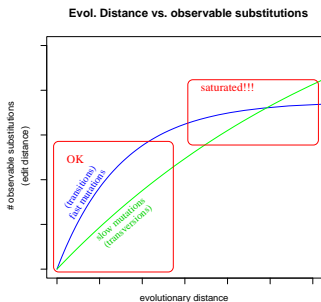
Saturation Plots can be created, e.g., with the DAMBE program as follows

- 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)

Saturation Plots for DNA in DAMBE

Saturation Plots can be created, e.g., with the DAMBE program 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.



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

The same can be done for amino acids

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

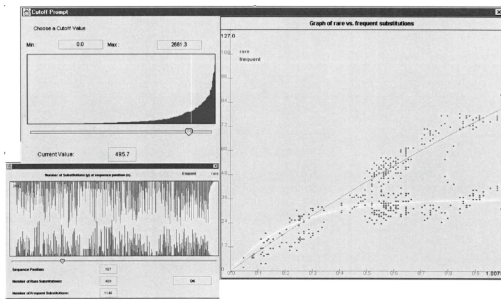
The same can be done for amino acids, but

- There is not intuitive division into fast and slow substitutions

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,
- AsaturaA 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 slow. (There is no guideline for that choice.)
- Then the numbers of fast and slow substitutions are plotted against the distance accordingly.



- We can compute a likelihood value for a tree based given an alignment and model... (more on that tomorrow).

Posterior Probabilities and Empirical Bayes

- We can compute a likelihood value for a tree based given an alignment and model... (more on that tomorrow).
- Problem: How different are the likelihoods?

Posterior Probabilities and Empirical Bayes

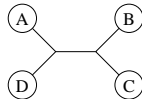
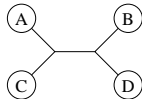
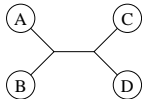
- We can compute a likelihood value for a tree based given an alignment and model... (more on that tomorrow).
- Problem: How different are the likelihoods?
Just from the value of likelihoods one often cannot tell whether they are significantly different.

Posterior Probabilities and Empirical Bayes

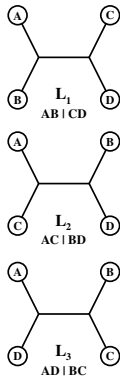
- We can compute a likelihood value for a tree based given an alignment and model... (more on that tomorrow).
- Problem: How different are the likelihoods?
Just from the value of likelihoods one often cannot tell whether they are significantly different.
- Normalization: Posterior probabilities are computed:

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

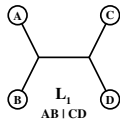
- We can use that on the three different quartet topologies to assess phylogenetic information in our data.



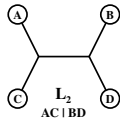
Plotting Posteriors: Likelihood Mapping



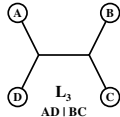
Plotting Posteriors: Likelihood Mapping



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

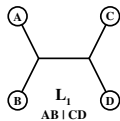


$$p_2 = \frac{L_2}{L_1 + L_2 + L_3}$$

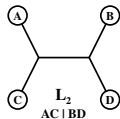


$$p_3 = \frac{L_3}{L_1 + L_2 + L_3}$$

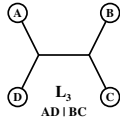
Plotting Posteriors: Likelihood Mapping



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



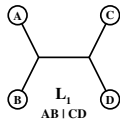
$$p_2 = \frac{L_2}{L_1 + L_2 + L_3}$$



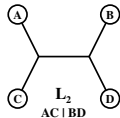
$$p_3 = \frac{L_3}{L_1 + L_2 + L_3}$$

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

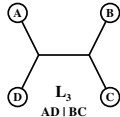
Plotting Posteriors: Likelihood Mapping



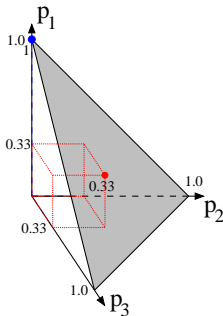
$$p_1 = \frac{L_1}{L_1 + L_2 + L_3}$$



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

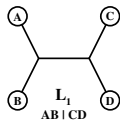


$$p_3 = \frac{L_3}{L_1 + L_2 + L_3}$$

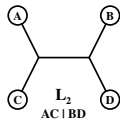


Since $p_1 + p_2 + p_3 = 1$, 3D points (p_1, p_2, p_3) fall into a triangular (simplex).

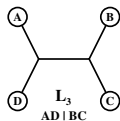
Plotting Posteriors: Likelihood Mapping



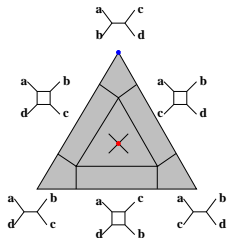
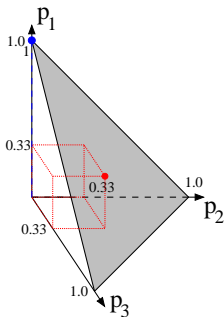
$$p_1 = \frac{L_1}{L_1 + L_2 + L_3}$$



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

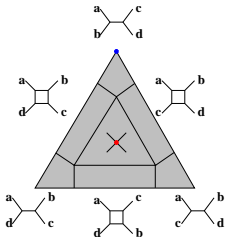
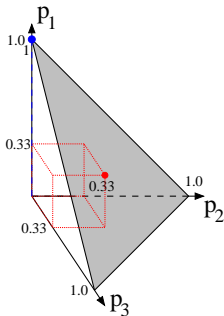
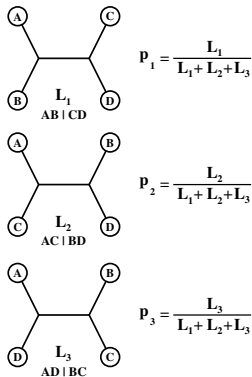


$$p_3 = \frac{L_3}{L_1 + L_2 + L_3}$$



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

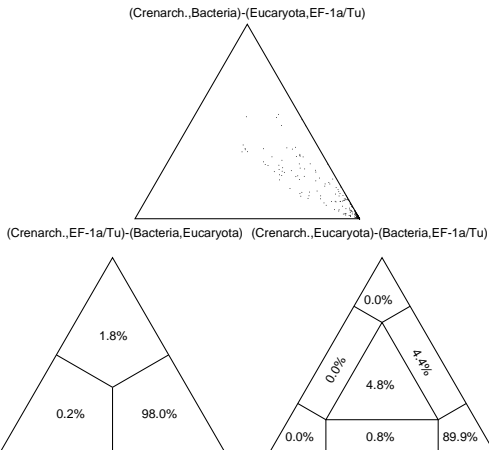
Plotting Posteriors: Likelihood Mapping



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.

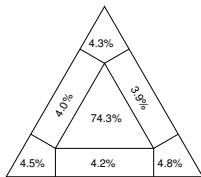
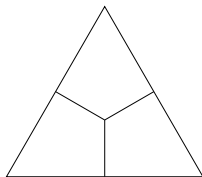
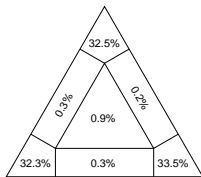
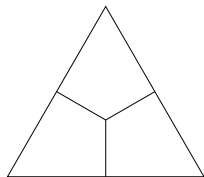
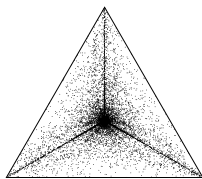
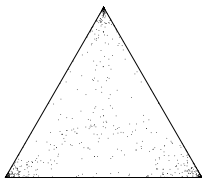
Likelihood Mapping (Cluster Analysis)



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.

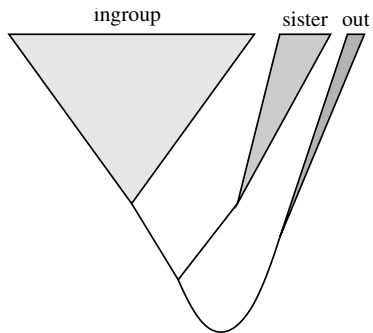
Likelihood Mapping (Information Content)



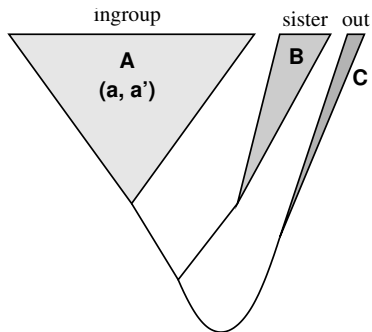
The Simplex Plot can also visualize the information content in an alignment.

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**.

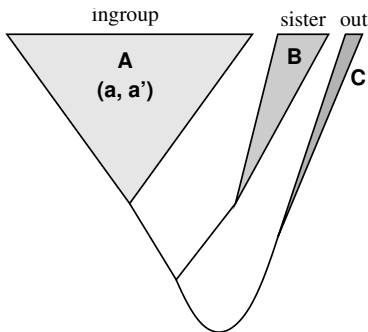
Likelihood Mapping to Validate Outgroups



Likelihood Mapping to Validate Outgroups

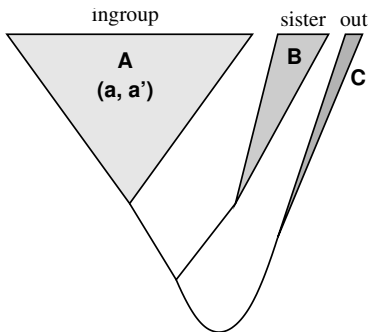


Likelihood Mapping to Validate Outgroups



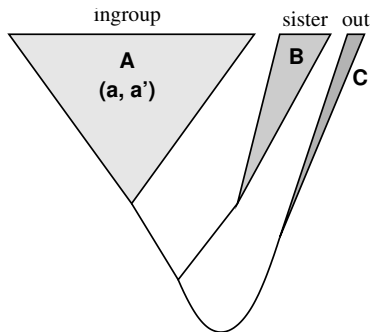
- We can check the reliability of an outgroup by assigning taxa to three sets: **A** - the examined ingroup, **B** - an early sister group, and **C** - the outgroup.

Likelihood Mapping to Validate Outgroups



- We can check the reliability of an outgroup by assigning taxa to three sets: **A** - the examined ingroup, **B** - an early sister group, and **C** - the outgroup.
- random quartets are drawn from the sets: two from **A** and one each from **B** and **C**.

Likelihood Mapping to Validate Outgroups



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