Higher Order Substitution Models - Mixture Models in Phylogenetics

Killian Smith [1,2] Sebastian Höhna [1,2]

[1] GeoBio-Center LMU, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 Munich, Germany
[2] Department of Earth and Environmental Sciences, Paleontology and Geobiology, Ludwig-Maximilians-Universität München, Richard-Wagner-Str. 10, 80333 Munich, Germany

Mixture Model

Mixture models in phylogenetics are a natural extension from **partitioned models**. In an analysis using **partitions**, different sites in the alignment are assigned to follow different substitution models. However, the researcher must predetermine which sites follow which model.

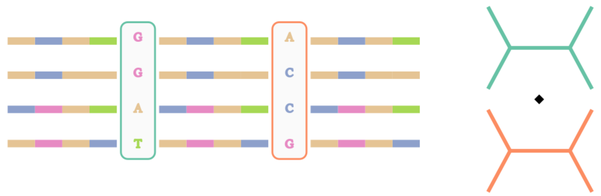


Figure 1. Partitioned Likelihood (products)

An analysis using **mixtures** does not have this limitation, allowing for an analysis that considers the posiblitity of all substitution models at all sites.

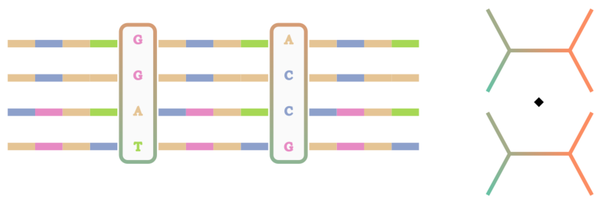
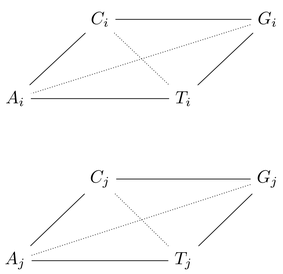


Figure 2. Mixture Likelihood (coproducts)

This is achieved by constructing a higher-order substitution model as the coproduct of the component substitution models.



$$Q_i + Q_j = \begin{pmatrix} Q_i & 0 \\ 0 & Q_j \end{pmatrix}$$

Figure 3. Mixture Model Markov Chain

Existing models that account for site heterogeneity include:

- **CAT**: Uses a Bayesian framework to infer the number of categories and their values.
- **C10-C60**: Use emperically generated categories.
- **EDCluster**: Use clustering to emperically determine categories.

In our work, we add support for inference of fixed sized categories in **RevBayes**, and analyze their effectiveness in a simulation study as well as look at empirical examples.

Simulation Study

We simulated 1000 DNA sites accross a predefined 6 taxon tree with 3 instances of short-long branch pairings using the *F81* model with 1, 2, 4, 8, and 16 categories.

We then attempted to reconstruct the trees from each of the simulated *MSA*s using **RevBayes** assuming the *F81* model with 1, 2, 4, 8, and 16 categories.

Each analysis was completed with 2 replicates, and each *simulated-assumed* analysis was repeated on 4 seperate simulations.

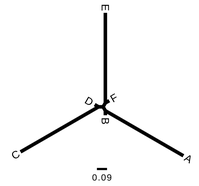


Figure 4. True Tree

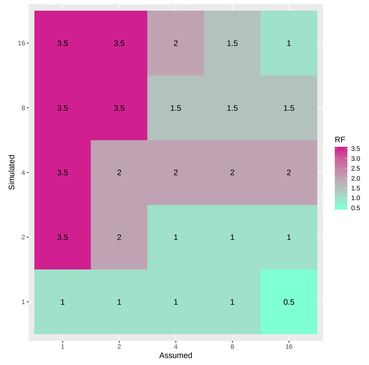


Figure 5. Average RF Distance from True Tree

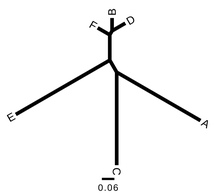


Figure 6. Under-Specification:
Simulate F81pi16 - Analysis with F81

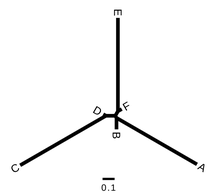


Figure 7. Over-Specification:
Simulate F81 - Analysis with F81pi16

Conclusions

From the simulation study, it is appearent that model under-specification can lead to long-branch-attraction (*LBA*) artifacts.

There does not appear to be any adverse effects as a result of model over-specification in regards to number of categories, and the true tree is still recovered.

The BearIRBP genes provides support that there does exist site heterogeneity in the real wold, even in DNA datasets.

Our general recomendation from this study would be to use a generous number of categories, pending computational resources.

Empirical Evidence

In addition to the simulation study, we also see some empirical evidence for mixture models in real world data. These tracer plots are from the *bears_irbp* gene dataset (available from *RevBayes* tutorial).

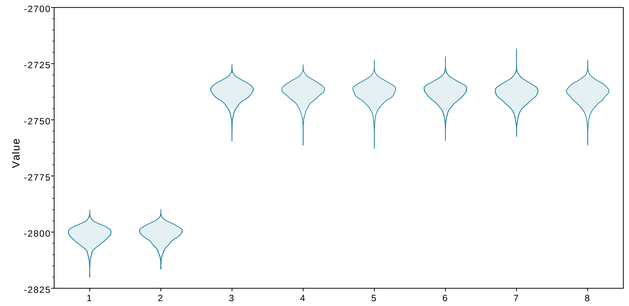


Figure 8. BearsIRBP Log-Likelihood under F81 (1,2,4,8 Cats : 2 reps)

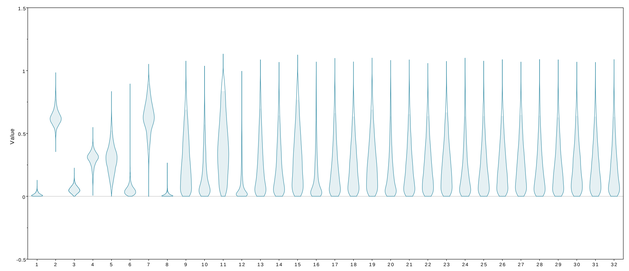


Figure 9. BearsIRBP Base Frequencies under F81 with 8 Categories

Future Work

The work presented here is on a small scale. We are working on a larger and more comprehensive simulation study.

In addition, we will also perform the same analysis for amino acid sequences.

We would also like to apply these methods to more real world datasets.

There is still more work to be done to improve mixing and convergence for models with a large number of categories.

Acknowledgments

Deutsche Forschungsgemeinschaft

Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities