PHYLOGENETICS WITH REAL CONFIDENCE

Steffen Klaere

s.klaere@auckland.ac.nz

THE UNIVERSITY OF AUCKLAND
FACULTY OF SCIENCE
Department of Statistics

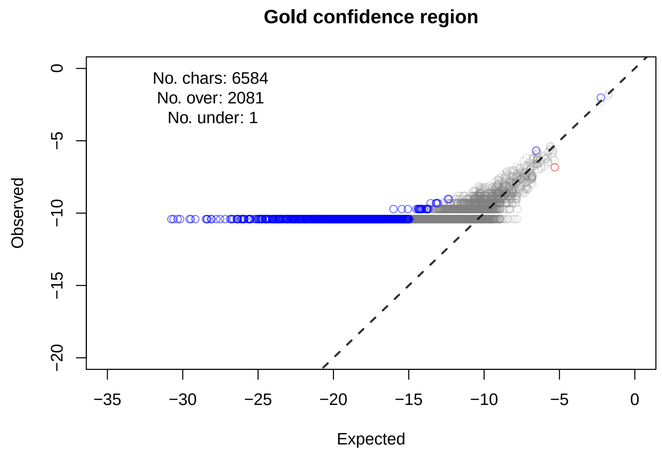
THE PROBLEM

The approach discussed here aims at statistically identifying those characters which are not well explained by an inferred phylogenetic model. This problem is part of the MISFITS approach [5] which attempts to identify the sites which are well explained by the model and the sites which are influenced by additional processes.

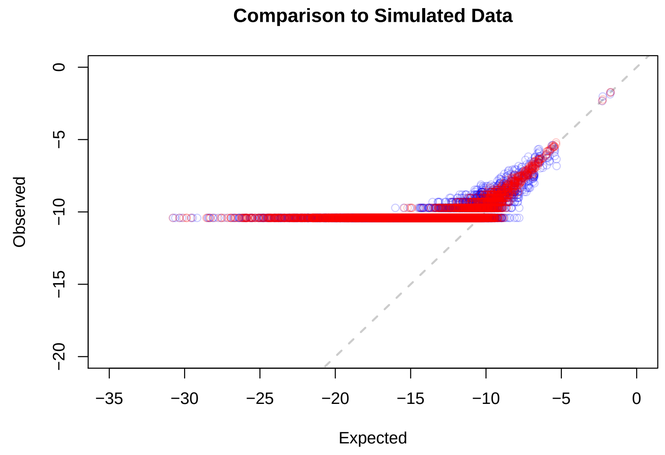
In our original work we used the Gold confidence region

$$CI_G = p \pm \sqrt{\chi^2_{\alpha,df}} \sqrt{\frac{p(1-p)}{n}},$$

where $\chi^2_{\alpha,df}$ is the α quantile of a χ^2 distribution with df degrees of freedom. This approach predominantly highlights single occurrence sites as insufficiently explained. This nicely conforms with methods of stability analysis such as SLOW-FAST [1], where sites with high parsimony score were highlighted and subsequently removed. In the following figure we plotted observed site frequency against the site likelihood under the chosen model (HKY+Γ with 8 rate categories). Coloured circles correspond to sites identified as insufficiently explained.



However, simulations under the model indicate that statistically single occurrence sites are not necessarily outliers of the model as the following plot shows.



This observation leads to two questions:

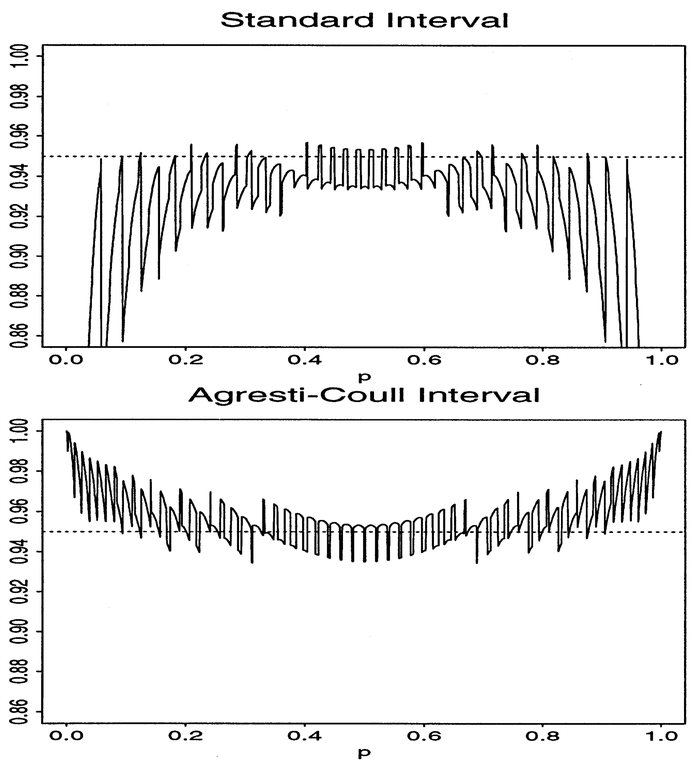
- What are statistically meaningful confidence regions for our data?
- Is the selection using confidence regions biologically meaningful?

Here, I will investigate the first question but am happy to discuss potential ideas about the second.

MULTINOMIAL AND BINOMIAL CONFIDENCE REGIONS

We address this problem by investigating common problems of categorical variables, and potential solutions to the problem.

One criterion to assess the quality of a confidence region is to look at its coverage probability, i.e. the proportion of times, the true parameter is in the confidence region. This has been identified as a problem for binomial as well as multinomial random variables [2, 3].



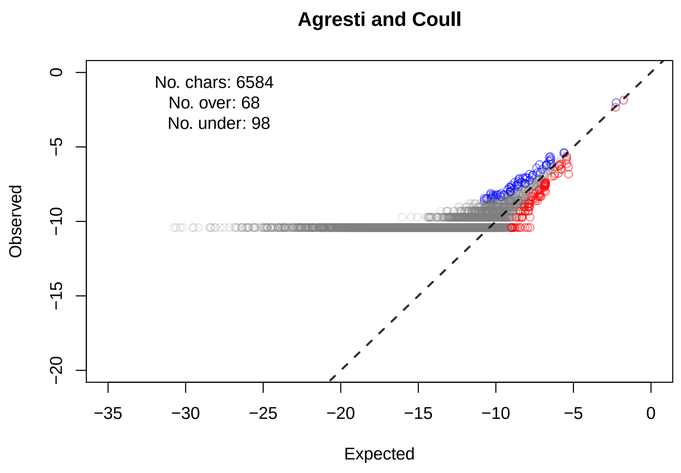
The figure shows the result for two confidence intervals for binomial proportions. We see that the standard interval, which is related to the Gold confidence interval covers particularly bad for small proportions, a property that most single occurrence sites share. Agresti-Coull is one of the suggestions made by [2] for a good coverage.

I decided to use the Agresti-Coull approach to compute an alternative confidence region for this problem with surprisingly positive results despite rather bluntly ignoring the multinomial nature of the sites. The approach is skewing the proportion parameter such that

$$\widetilde{p} = \frac{np + q^2_\alpha/2}{n + q_\alpha},$$

$$CI_{AC} = \widetilde{p} \pm \alpha \sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n}},$$

where q_α is the quantile of the standard normal distribution for confidence level α. Using this confidence region on our problem yields the following graph:



i.e. a much better fit to our observations from the simulations. [2] also suggest the Clopper-Pearson region which computes the bounds as quantiles of a β distribution whose parameters depend on the observed values.

While this approach seemingly addresses the problem of having confidence regions that address the similarity between data and simulation, it is necessary to find methods that address the inherent categorical nature of the data. [4] provides an overview of confidence regions for multinomial distributions based on asymptotic properties. However, most of these properties do not hold for the sparse categorical data we deal with.

[3] points out, that usually confidence regions either minimise volume or maximise coverage, with the Gold confidence region belonging to the former, and the Clopper-Pearson confidence region belonging to the latter. [3] also suggest an alternative confidence region, incorporating the Clopper-Pearson approach, which turns out to find a reasonable compromise. However, their confidence region is one-sided and thus only of limited use to us.

A further downside of multinomial confidence regions is the simultaneous calculation which makes it hard to identify sites insufficiently explained by the model, a property I consider essential. It might be more fruitful to investigate the nature of the graphs from the regression point of view.

REFERENCES

References

[1] Brinkmann H, Philippe H. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol Biol Evol* 16(6):817–825, 1999.

[2] Brown LD, Cai TT, DasGupta A. Interval Estimation for a Binomial Proportion. *Stat Sci* 16(2):101-133, 2001.

[3] Chafaï D, Concordet D. Confidence regions for the multinomial parameter with small sample size. *J Am Stat Assoc* 104(487):1071–1079, 2009.

[4] Lee AJ, Nyangoma SO, GAF Seber. Confidence regions for multinomial parameters. *Comput Stat Data Ana* 39:329–342, 2002.

[5] Nguyen MAT, Klaere S, von Haeseler A. MISFITS: Evaluation the Goodness of Fit between a Phylogenetic Model and an Alignment. *Mol Biol Evol* 28(1):143–52, 2011.