Variance reduction and uncertainty quantification for phylogenetic MCMC

Dr Jere Koskela (Statistics) and Dr Jerome Kelleher (Oxford Big Data Institute)

Phylogenetic tree inference from DNA sequence data is a central technique in the study of pathogens. Programs such as BEAST [6], which use MCMC inference, have been cited in thousands of studies, and most recently, have been vital to our understanding of SARS-CoV-2. BEAST estimates a posterior distribution of phylogenetic trees and model parameters that are consistent with observed DNA sequence data, and this distribution takes the form of a set of (trees, parameters) tuples sampled at regular intervals from the Markov chain. This process of taking a subset of the Markov chain (or, put another way, of discarding most of the generated samples) is known as "thinning", and is necessary because of the computational burden of storing and processing a large number of trees. However, thinning has been shown to reduce the precision of MCMC inference, and always produces worse results than an unthinned chain [1] unless the cost of storing and processing samples is non-negligible [4]. As the size of datasets becomes larger (e.g. hundreds of thousands of SARS-CoV-2 sequences are now available), thinning becomes more and more severe because of computational constraints, and the effect of this on the accuracy of phylogenetic inference is not well understood.

A recent breakthrough in computational genomics may provide an answer to the problem, by using a novel technique for genealogical tree compression. The "succinct tree sequence" data structure [3] is a method of storing a large collection of correlated trees, currently used to efficiently represent the ancestral history of recombining organisms. It has so far lead to performance improvements of multiple orders of magnitude in genome simulation [2], computation of summary statistics from trees [5], and ancestry inference [3]. There is great potential for the application of this technique to Bayesian phylogenetics, for example by enabling the design of efficient thinning schedules, rather than being forced to thin extensively by computational budget constraints.

Proposed research plan:

1. Become familiar with the coalescent model, and MCMC methods for likelihood-based inference under the coalescent (no prior knowledge is required).
2. Become familiar with the relationship between the autocorrelation of a Markov chain, the per-sample cost of storing and processing states, and the effect of thinning on algorithmic efficiency.
3. Derive the asymptotic computational cost of appending a tree into a succinct tree sequence, and of evaluating standard summary statistics on trees stored as a succinct tree sequence.
4. Adapt an existing MCMC tree inference method to store the output chain as a succinct tree sequence using the tskit C library.
5. Estimate the autocorrelation from algorithm runs for simulated and/or public data sets covering a range of sample sizes.
6. Time permitting, develop predictions of good thinning schedules for large sample sizes based on 3. and 5.

Note that this project contains a substantial software development component and requires a good working knowledge of Python.

References

[1] C G Geyer. Practical Markov chain Monte Carlo. Statistics and Computing 7(4):473–483, 1992.

[2] J Kelleher, A M Etheridge, and G McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Computational Biology 12(5):e1004842, 2016.

[3] J Kelleher, Y Wong, A W Wohns, C Fadil, P K Albers, and G McVean. Inferring whole-genome histories in large population datasets. Nature Genetics 51:1330–1338, 2019.

[4] A B Owen. Statistically efficient thinning of a Markov chain sampler. Journal of Computational and Graphical Statistics 26(3):738–744, 2017.

[5] P Ralph, K Thorton, and J Kelleher. Efficiently summarizing relationships in large samples: a general duality between statistics of genealogies and genomes. Genetics 215(3):779–797, 2020.

[6] M A Suchard, P Lemey, G Baele, D L Ayres, A J Drummond, and A Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution 4(1):vey016, 2018.