Tree-sequence methods in Genealogical analysis of large whole-genome datasets

CMEE MRes project

Wenhua Zhou

wz2812@ic.ac.uk

Supervisor: Professor Tim Barraclough

Co-supervisor: Dr. James Rosindell

Keywords: Tree sequences, Phylogenetics, Large datasets, Genomes, Genealogical analysis, Aspergillus

Department of Life Sciences
Imperial College London
United Kingdom
December 9, 2019

1 Introduction

Genealogical analysis of DNA sequences is an increasingly popular task in evolutionary biology. There has been massive growth in the availability of whole-genome data, which brings the potential to infer useful information about the traits of this species. However, current methods performing accurate genealogical analysis are not computational efficient enough to be used on large datasets with many individuals and species. Recently, a tree-sequence recording method (Kelleher et al., 2016) was introduced for large scale genetic data. This approach represents genome data in the form of correlated trees using efficient coalescent simulation. By this method, it manages to increase computational efficiency dramatically. We will adjust and implement the similar tree-sequence method into *Aspergillus* whole-genome dataset from Imperial's Department of Infectious Disease and Epidemiology.

2 Methodology

Compared to the ancestral recombination graphs (Griffiths and Marjoram, 1996), tree sequences (Kelleher et al., 2019) encodes a sequence of marginal trees instead of common ancestors and recombination events. Conventional data storage of a genomic variation dataset is a matrix, requiring sample size n rows and site m columns which need $O(mn)$ space. Instead, tree sequences represent the data as a marginal tree, which can be fully encoded by samples and ancestors with related mutation sites using only $O(n + m)$ space to store. An algorithm was developed (Kelleher et al., 2016) to generate tree sequences and count sample size from stored data.

A whole-genome dataset consists of millions of base pairs, and there are millions of sites per chromosome. Using tree sequences reduces the space of storing whole-genome dataset significantly, and this makes further genealogical analysis feasible. Additionally, tree-sequence data also provide comparable accuracy to ARGs and other current methods.

However, original whole-genome data contains DNA sequences of million SNPs instead of genealogical trees. A pre-processing step with appropriate algorithms is therefore required to transform DNA sequences into trees. A huge amount of methods are developed to infer trees, distance methods and the principle of parsimony are the two most commonly used methods. Distance methods determine the trees by calculating the distance between DNA sequences, this can either be the proportion of differences or derived from sequence models. The principle of parsimony simply chooses the tree with the least number of mutations and recombination events after counting up the number of changes in each tree. Another common approach is by maximum likelihood analysis (Yang, 2007), it generates a new tree and keeps rearranging itself until it optimizes the likelihood.

3 Future work

There are a few other methods for generating trees such as Bayesian models using MCMC. These models generally provide solid approximations of trees but also requires more computational resources. When applying on *Aspergillus* whole-genome data, comparison in different methods is essential to provide a suitable model with comparable accuracy and computational efficiency.

After getting the whole-genome data by tree sequence methods, I will introduce evolution models on the encoded data to show the possibility of further genealogical analysis. I will also compute results and compare with previously existing methods. Using the results of those comparisons, I will discuss the strength and limitations of tree-sequence methods, and propose a few ideas to improve the method and models.

4 Project Timeline

In this 9-month project, no budgets are required. The whole project can be separated into 5 steps:

- **Preparation**: Do essential reading for the project, contact with people from Imperial's Medical department providing data and from University of Oxford researching on tree-sequence methods. This is planned to finish by the start of Spring term.
- **Pre-processing**: Design algorithms to transform genome data into tree sequences. This step takes a few weeks and I plan to finish it by January.
- **Tree-sequence methods**: This is the main target for the project and is expected to take 4 months from February to May.
- **Comparing results**: Compare the results of tree-sequence methods and explore how this methods make further genealogical analysis feasible. This is planned to finish by June.
- **Conclusion, thesis writing and final check**: For the last 2 months of the project, I will discuss the results and make conclusion for tree-sequence methods. And then I will also check all the contents and write them into the MRes thesis.

Bibliography

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology*, 3(4):479–502, 1996.

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.

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(9):1330–1338, 2019.

Z. Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8): 1586–1591, 2007.

Supervisor signature:

I have seen and approved the proposal and the budget.