First Semester Report

Lénárd Lajos Szánthó (lenard.szantho@ttk.elte.hu)
Doctoral School of Physics
Statistical Physics, Biological Physics
and Physics of Quantum Systems Program
Supervisor: Gergely J. Szöllősi

Thesis title: Developing next-generation phylogenetic methods

January 21, 2022

Introduction

Phylogenetics studies the evolutionary relationships between organisms by reconstructing their tree of descendants, called a phylogeny. The tree inference process is a complex, hierarchical and to some extent modular statistical machinery which developed in parallel with the advances in gene and genome sequencing and the availability of increasing computational resources. Better statistical models, more sequence data and stronger computational power results in theory in better estimates of the phylogeny. However, these models only consider evolution at the level of genes, resulting in reconstructing gene trees and not the species tree. Single genes can have considerably different, sometimes contradictory histories, yet we know that they evolved along a common tree, the tree defined by the species they reside in today (Maddison, 1997). The current praxis of removing contradictory genes from the analyses and only considering homologous genes is neglecting high amounts of data that – given appropriate models – could help resolve questions left undecided in the best case, or inferred wrongly in the worst case by the conventional models. Advanced models, such as ALE (amalgamated likelihood estimation) (Szöllősi et al., 2013) were developed to account for evolutionary events which can be defined at the level of species tree: gene loss, gene duplication and horizontal gene transfer. These new models allow a species tree aware phylogenetic reconstruction (gene tree-species tree reconciliation). During my PhD I will build on this base to develop new phylogenetic methods in order to answer questions still under debate in the scientific community.

The currently planned or ongoing projects in this regard are:

- **New methods for deep phylogenies – modelling the eukaryogenesis**
  One major evolutionary transition was the emergence of eukaryotes by an endosymbiosis between an archaeon and a bacterium between 2.5-2 billion years ago. Many aspects of this event however, are not yet understood.

A project realised in collaboration with Phylogenecisists and Microbiologists from the University of Bristol, UK and the NIAZ, The Netherlands. The subtasks connected with this project are:

  - **Novel clustering methods**
    Going back billions of years based on the genomes of extant species is challenging because novel mutations may overwrite old ones. Clustering methods, which form one of the crucial steps of phylogenetic inference, may not detect relationships between such genes with efficient accuracy and sensitivity. A novel approach would be a hierarchical, tree-aware clustering method, which starts from a conventional phylogenetic clustering but then the gene trees would be used to reconstruct their common ancestor's ancestral sequence and run a second clustering procedure on these ancestral sequences. The new clusters would be translated back to the old ones by merging some of those which seemed distinct based on the extant sequences but are more similar considering the ancestral sequences. This method could be applied iteratively.
  - **Horizontal gene transfer highways**
    Horizontal gene transfers (HGTs) contribute about 30% to the evolution of bacteria, as recently shown in a study by Coleman et al. (2021) in which I was also involved. HGT means that a gene can be shared between species even if they are distantly related to each other. An endosymbiosis event can be imagined as a massive correlated HGT event, where not just one gene but a considerable part of a whole genome gets integrated into the genome of another life form, the archaeon, in the case of Eukaryotes. The merging of genomes would leave a specific fingerprint on the trees: the change in the rate of gene duplication, transfer and loss. The task here is to extend ALE to model the change of these rates, and thus be able to detect such massive correlated HGT events. This would be a new endosymbiosis-aware gene tree-species tree reconciliation method.

- **New method to detect and ameliorate long branch attraction (LBA) artefacts: CAT-PMSF**
  This project aims to provide a new method called CAT-PMSF to detect and ameliorate long branch attraction (LBA) artefacts, a well-known systematic bias in phylogenetic inferences leading to debates about, e.g. whether the first animal was a sponge or a jellyfish. The CAT-PMSF pipeline combines existing phylogenetic software tools and methods to provide the framework to assess any custom dataset in an easy and fast manner. Inferences that required months of runtimes now can be made in several days with CAT-PMSF, and it gives consistent results.

- **Date the tree of fungi**
  A project in collaboration with László Nagy and his research group at the Biological Research Centre (BRC) at Szeged aims to create a dated tree of fungi. The kingdom fungi consists of single- and multicellular eukaryotes and shows excess amoount of horizontal gene transfers (HGT) despite the separated genome storing

and expression environment (i.e. has a nucleus). For this reason, applying species tree aware reconstruction should help resolve the more than 1 billion-year-old history of fungi. First, we have to build good quality gene trees and have a prior on the species tree. We can then reconcile using ALE to yield the final species tree accounting for gene duplications, losses and transfers. This tree can then be used with fossils to trace back the most likely range of age of the internal nodes.

- **Date the tree of bacteria**
  A continuation of the rooting the tree of bacteria project published in 2021 Coleman et al. (2021) is to extend the dataset and date the tree of bacteria.
  A project realised in collaboration with Phylogenecisists and Microbiologists from the University of Bristol, UK, the NIAZ, The Netherlands and The University of Queensland, Australia.

Description of research work carried out in current semester

In this semester the main focus was on finishing the project *New method to detect and ameliorate long branch attraction (LBA) artefacts: CAT-PMSF* and summarize the results in a paper. The manuscript is 99% done, finishing steps and submission can be continued once the exam period ends.

Secondary focus was given to the implementation of *Horizontal gene transfer highways* for which a proof of concept code and simulations were written in Python and the current state was presented on a conference.

Tertiary focus was granted for continuing the *Date the tree of bacteria* project, MCMC analyses were run on a 1000-taxa concatenate of 71 genes, sadly it did not converge after approximately 150 days despite the lots of allocated resources, so the subsampling will be inevitable.

For the *Date the tree of fungi* project the to species to be included in the analyses were selected, obtained and preliminary analyses of BLAST and clustering were performed. To be continued in the next semester.

To prepare for the *Novel clustering methods* project I have enrolled to the *Clustering with networks* course.

Publications

We plan to submit our manuscript summarizing the project *New method to detect and ameliorate long branch attraction (LBA) artefacts: CAT-PMSF* to the Molecular Biology & Evolution (MBE) in the following weeks.

Studies in current semester

I have enrolled to the following classes:

INFPHD412-N Bioinformatics (Vince Grolmusz)

FIZ/3/010E Sensory biophysics (Gábor Horváth)

FIZ/3/064E Clustering with networks (Gergely Palla, Péter Pollner)

The Bioinformatics course was graded to 5, the other two courses are still under completion when writing this report.

Conferences in current semester

I have attended the *Moore-Simons Project on the Origin of the Eukaryotic Cell annual meeting* (October 25-26, 2021) and presented a poster about the progress on the Eukaryogenesis project and about the implementation of highway transfers in the ALE-framework. The conference was held online due to the COVID19 pandemic.

Professional activities

I have presented a lecture on the Night of the Researchers (Kutatók Éjszakája, September 24, 2021) with the topic: ,,Harc a bitekkel, avagy a nagy számítási teljesítményű (HPC) klaszterek kihívásai a kutatásban".

I have helped in the development of the Kooplex Research and Teaching System, which provides on-demand JupyterNotebook access to the students and researchers of the university (or collaborators of). My task was the refinement of the central authentication system. We will continue improving the system and implementing new features in the following semester(s).

I have maintained the research group's high performance computing (HPC) cluster, one node with 1 TiB of RAM was created to support further investigations on the project *Date the tree of bacteria*.

I have helped to maintain the university's high performance computing (HPC) cluster which service is provided by the ELTE IIG.

References

Coleman, G. A., Davín, A. A., Mahendrarajah, T. A., **Szánthó, Lénárd L.**, Spang, A., Hugenholtz, P., Szöllősi, G. J., and Williams, T. A. (2021). A rooted phylogeny resolves early bacterial evolution. *Science*, 372(6542):eabe0511.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523.

Szöllősi, G. J., Roskiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912.