Fourth 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

June 6, 2023

Introduction

Phylogenetics studies the evolutionary relationships between organisms by reconstructing their tree of descendants, called a phylogeny. During my doctoral studies, I am developing new methods and applications as described in the first semester's report in the following main projects:

- New methods for deep phylogenies – modelling the eukaryogenesis
- New method to detect and ameliorate long branch attraction (LBA) artefacts: CAT-PMSF
- Date the tree of fungi
- Date the tree of bacteria

Description of research work carried out in the past three semesters

2021/22 fall semester

In the 2021/22 fall 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.

2021/22 spring semester

In the beginning of the 2021/22 spring semester the manuscript describing our results in the project *New method to detect and ameliorate long branch attraction (LBA) artefacts: CAT-PMSF* was finished and submitted to the journal Systematic Biology. The review process has begun and early June we got positive feedback: *accepted with minor revision.* Currently I am revising the article and hope to submit it by middle of July.

The implementation of *Horizontal gene transfer highways* was continued, currently there are two tools written in Python, one simulating the evolution in forward direction and another (the original one) tracing it back. This allows us to test the predictions against simulations. Further work was devoted to implement the Conditional Clade Probabilities, which allows the software to consider many tree topologies with their appropriate probability weights during estimating the likelihood of the gene trees given the species tree.

The *Date the tree of fungi* project preliminary trees were obtained which had bogus topology (some species of the taxon Zoopagomycota were clustering with another taxon Mucoromycota), we were cheking whether the sequence data may have been contaminated, but it is not, more sophisticated model needs to be used to describe the data.

2022/23 fall semester

During the 2022/23 fall semester the manuscript describing our results in the project *New method to detect and ameliorate long branch attraction (LBA) artefacts: CAT-PMSF* was going through a second round of revision (first revision was sent in in August, second round of reviews were received in November, and the second revision was sent in in December) at the journal Systematic Biology. The review process has prolonged more than we were expecting, but has greatly imporved our manuscript and as of writing this report we are waiting for the decision of the Editor in Chief.

The implementation of *Horizontal gene transfer highways* was continued, the code was refractored in Python, various unit tests were made, but unfortunately due to personal reasons no signifcant progress could be made after November.

The *Date the tree of fungi* project has moved on during the semester with great pace, we were able to determine a dated tree using fossil calibrations and using McmcDate (`https://github.com/dschrempf/mcmc-date` developed by one of our research group members, Dominik Schrempf. Having only two maximum calibrations has made the inference less conclusive as we were hoping for and during investigating the results with our collaborators (see Conferences section), one of the maximum calibrations prooved to be unreliable which may turn this project to completely new tracks and an exciting publication is to be expected.

Connected to the *Date the tree of bacteria* project, a manuscript was prepared with our collaborators (involved in the original paper Coleman et al. (2021)) explaining the methods and controversies when working with deep phylogenies. Expected to be published during 2023.

Description of research work carried out in the fourth semester

During this semester the manuscript describing our results in the project *New method to detect and ameliorate long branch attraction (LBA) artefacts: CAT-PMSF* was accepted and published on 22nd of March 2023 by the journal Systematic Biology (see Publications section).

For the *Date the tree of fungi* project new analyses were made using the softwares ALE (Szöllősi et al., 2013) and Notung (Durand et al., 2006) to asses the effect of transfers on the number of gene copies at the root of eukaryotes required to explain the evolutionary history of pectinase genes. Using updated calibrations a new dated tree analsys was run. With the new results we can start writing the manuscript and submit it in the next 1-2 months.

Connected to the *Date the tree of bacteria* project, a manuscript was submitted with our collaborators (involved in the original paper Coleman et al. (2021)) explaining the methods and controversies when working with deep phylogenies. We got rejected and the reviewers have proposed to restructure the paper focusing on only one train of thought. The new version has just been submitted to the journal Genome Biology and Evolution.

Another collaboration paper as a spin-off of the *Date the tree of bacteria* project was submitted to Nature Microbiology with the aim to shed light on the origin (and time of origin) of the ATP synthese genes (see Publications section), we are waiting for the reviews.

During this semester a new project was started assessing the capability of Machine Learning Protein Language models applied on gene sequences to predict phylogenetically and evolutionally relevant quantities (i.e. per site amino acid frequencies). If we succeed the results could be used to speed up the CAT-PMSF method even more and would open the door to further exciting projects.

One of my long-term personal goals is to apply my knowledge about phylogenetics in research focusing on the gut microbiome. Recently we have found and made contact with a potential collaboration partner located in Christian-Albrechts-Universität zu Kiel, Germany. I plan to apply for the EMBO Scientific Exchange Grant to start the collaboration and travel there in the next semester.

Publications

Published

- **Lénárd L Szánthó**, Nicolas Lartillot, Gergely J Szöllősi, Dominik Schrempf (2023). Compositionally Constrained Sites Drive Long-Branch Attraction. *Systematic Biology*, 10.1093/sysbio/syad013. IF: 9.16

- 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. IF: 63.832, cited: 87 (Google Scholar)

Submitted (preprint available)

- Tom A. Williams, Adrian A. Davin, Benoit Morel, **Lénárd L. Szánthó**, Anja Spang, Alexandros Stamatakis, Philip Hugenholtz, Gergely J. Szöllősi. The power and limitations of species tree-aware phylogenetics. `https://doi.org/10.1101/2023.03.17.533068`

- Tara A. Mahendrarajah, Edmund R. R. Moody, Dominik Schrempf, **Lénárd L. Szánthó**, Nina Dombrowski, Adrián A. Davín, Davide Pisani, Philip C. J. Donoghue, Gergely J. Szöllősi, Tom A. Williams, Anja Spang. ATP synthase evolution on a cross-braced dated tree of life. `https://doi.org/10.1101/2023.04.11.536006`

Studies in current semester and before

2022/23 spring semester

FIZ/3/075E Extremes, Records, and Order-Statistics in Nature (Rácz Zoltán Attila), grade: 5

2022/23 fall semester

FIZ/3/017E Physics of environmental flows (Jánosi Imre Miklós), grade: 5

BIO/02/03E Human ethology (Kubinyi Enikő Dr.), grade: 5

2021/22 spring semester

BIO/02/01E Behaviour genetics (Enikő Kubinyi), grade: 5

BIO/10/32G Computer Modelling in Biology (Viktor Müller), grade: 5

2021/22 fall semester

INFPHD412-N Bioinformatics (Vince Grolmusz), grade: 5

FIZ/3/010E Sensory biophysics (Gábor Horváth), grade: 5

FIZ/3/064E Clustering with networks (Gergely Palla, Péter Pollner), grade: 4

Conferences during the four semesters

2022/23 spring semester

I will be attending the 29th National Meeting (28-31 August, 2023) of the Hungarian Biophysical Society (MBFT) and presenting a poster (or talk).

2022/23 fall semester

Moore-Simons Project on the Origin of the Eukaryotic Cell Annual Meeting
10-12 October 2022
attending

Ecological Days Conference 13-14 October 2023
presenting 10 minutes talk about the paper "A rooted phylogeny resolves early bacterial evolution" Coleman et al. (2021)

2021/22 spring semester

I did not attend any conference this semeseter.

2021/22 fall semester

Moore-Simons Project on the Origin of the Eukaryotic Cell Annual Meeting
25-26 October 2021
presenting poster

Teaching activities during the four semesters

2022/23 spring semester

I have thaught 1 lesson for the course *Statistical Physics B, elmfiz4bf19va* for BSc students.

2022/23 fall semester

No teaching activities this semeseter.

2021/22 spring semester

I thaught 3 computational practical classes for the course *Statistical Physics B, elmfiz4bf19va* for BSc students.

2021/22 fall semester

No teaching activities this semeseter.

Professional activities during the four semesters

in all four semesters

research group's HPC cluster development, maintenance, user support
44 compute node, performance: 57 TFLOPS FP64

ELTE Atlasz HPC cluster development, maintenance, user support
18 compute node, performance: 27 TFLOPS FP64

2021/22 fall and 2021/22 spring semester

Kooplex Research and Teaching System development

2021/22 fall semester

Night of the Researchers (Kutatók Éjszakája, September 24, 2021) lecture with the title: ,,Harc a bitekkel, avagy a nagy számítási teljesítményű (HPC) klaszterek kihívásai a kutatásban"

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.

Durand, D., Halldorsson, B. V., and Vernot, B. (2006). A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology*, 13(2):320–335.

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.