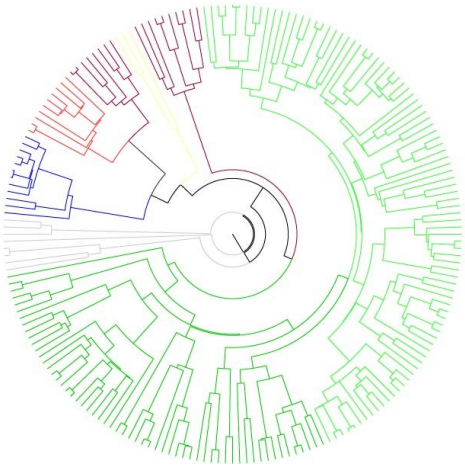
Phyloinformatics and a Pipeline for Construction of Supermatrices (PifCoSm)

Project work

This project will investigate different approaches to construct phylogenies from GenBank sequences. With its more than 200 million sequences, GenBank is the largest public repository for DNA sequences. It is therefore a good starting point when constructing phylogenies. However, it is not straightforward to make multi-gene phylogenies from GenBank data due to the non standard way and the quality of the annotations of the sequence metadata, and that the data has been collected for various disparate reasons and therefore is very heterogeneous. In this project you will test different ways to deal with these issues to produce reliable phylogenies. You will use an in-house pipeline (PifCoSm) to construct the phylogenies and test the pipeline.

You will work with real world data to reconstruct a phylogeny for a clade constituting two families of mushroom forming fungi, the Crepidotaceae and the Inocybaceae. The clade has been chosen since it is large enough to include many of the general problems using GenBank data and it is worthwhile automating the process, but it is small enough to construct phylogenies in a reasonable time frame such that many different approaches can be tested. There are several previous studies of the group so it is possible to evaluate if the produced phylogenies are reasonable, but there are also aspects that have not previously been resolved, such that the generated results will be interesting in themselves. During the project you will be part of the Ryberg research group at the Systematic Biology program, Evolutionary Biology Center, Uppsala University. The work of the project is expected to be included in a scientific publication. The project is suitable for a 10 to 15c project work.



You should:
Know how to reconstruct and interpret phylogenies including working with alignments, and be familiar with working on the command line and working with GenBank and using BLAST.

You will learn:
About challenges and potential issues working with multi-gene datasets and large phylogenies, about issues working with public sequences, how to handle potential taxonomic and nomenclature problems, hmmer and gene recognition, automation of phylogenetic reconstruction, and development of pipelines.

Possible extensions of the project:
For a person experienced in programming (Perl) it will also be possible to contribute to the development of the pipeline to adapt it to protein sequences, phylogenomics, gene tree/species tree methods, and/or other approaches. The gene recognition in the pipeline have to be adapted to gene region and taxonomic group, one way to expand the project is therefor to adapt and test it on other groups.

Contact:
Martin Ryberg, email: martin.ryberg@ebc.uu.se, tel.: 018 471 26 47
Systematic Biology, Department of Organismal Biology, Evolutionary Biology Center, Uppsala University