Bristol Bioinformatics Workshop: Bonus practical

Dr. Tom Williams

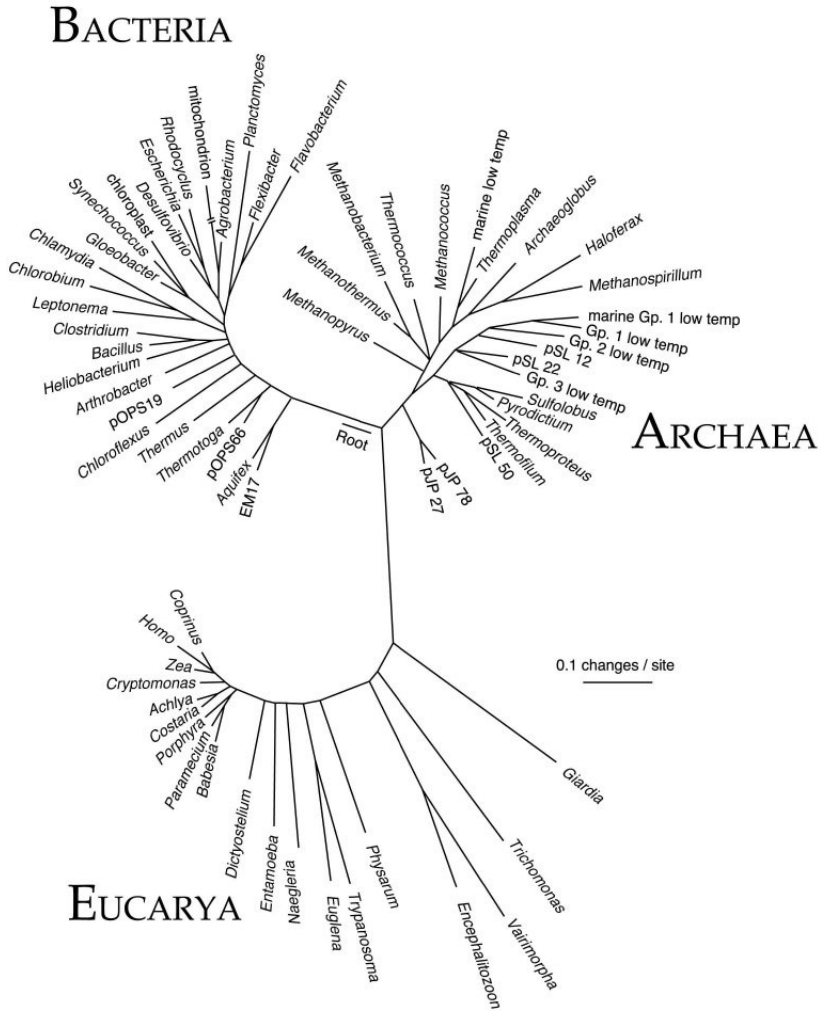
November 5, 2018

Background

This practical is for those who are thirsty for more after completing the "Florida Dentist" practical. It is somewhat more involved that the Dentist case, and involves (optionally) running some analyses on your own computer/compute resources, such as the Bayesian MCMC sampler PhyloBayes.

Phylogenetic analyses involve a series of decisions about methods and models, any of which may affect the outcome. We will explore the impact of these decisions on the phylogenies obtained, and the biological interpretations that result; for in-depth background on these issues, we recommend reading the book chapter (Williams and Heaps 2014). We will work with the famous "three domains" tree of life, and investigate the impact of taxon sampling and model choice on phylogenetic inference.

Building the tree of life



http://pacelab.colorado.edu/

The textbook tree of life is the "three domains" tree, in which bacteria, archaea and eukaryotes represent three separate domains. But these deep splits in the history of life occurred billions of years ago, and in reality we are not as sure about this tree as you might think. Ribosomal RNA is the molecule of choice for these cross-domain phylogenies, because it evolves very slowly and is found in all cellular lifeforms. Many, but not all, phylogenetic analyses of ribosomal RNA produce a "three domains" tree.

We will explore the effect of **alignment editing**, **model selection** and **taxon sampling** on trees of life inferred from ribosomal RNA genes. **How strong is the evidence for the "three domains" tree?**

Pace NR (1997) A molecular view of microbial diversity and the biosphere.

Science 276: 734-740.

Deliverables

Download the SSU and SSU+Tak datasets from

http://datadryad.org/resource/doi:10.5061/dryad.0hd1s.

You will need to align them, compare the fit of different evolutionary models, and build maximum likelihood and Bayesian trees using some of the modern software packages like IQ-Tree and PhyloBayes. The theory behind these different approaches will be discussed during the practical, as well as in depth in the accompanying book chapter.

In this practical, you will analyse two datasets: SSU and SSU+Tak. These are sets of small subunit (SSU) rRNA sequences from a selection of Bacteria, Archaea and eukaryotes. Compared to bacteria and eukaryotes, very few archaeal genomes have been sequenced. Recently, several new archaea have been discovered which are quite distantly related to the previously characterized groups: these include Thaumarchaeota, Aigarchaeota and Korarchaeota (TAK), which are added to the second dataset to recapitulate improvements in taxon sampling over the last few years.

1. Align the SSU dataset.
2. Infer a bootstrapped (1000 rapid bootstraps), maximum likelihood tree for this dataset using the GTR model with IQ-Tree. Compare the topology to the rRNA tree on the previous page.
3. Later, some more environmental Archaea were sequenced (found in the SSU+Tak dataset). Align and analyze the expanded dataset as before. Do you get the same result, with reference to the tree of life?
4. Use PhyloBayes to infer trees using both the GTR model and the CAT+GTR model (this step may take a few hours to a few days to run. Make sure to use the -s option to generate simulation files for the next step if using the original version of PhyloBayes; this is the default in PhyloBayes-MPI.
5. Compare the topologies obtained under the two different models. Are they the same or different? What predictions does each make about the relationship of the eukaryotes to other life forms?
6. Compare the fit of the GTR model with that of CAT using posterior predictive simulations - which is better? Which tree do you think is more reliable?
7. Compare the results obtained using the SSU and the SSU+Tak datasets. Briefly comment on the effect of taxon sampling and model selection on the trees obtained. How reliable, in your opinion, is the “three domains” tree?