**Ancient Sulfur Cycle Methods**

**Katherine Mateos**

**Summer 2020**

Many of the scripts reported here can be found at <https://github.com/carleton-spacehogs/sulfur>

**Constructing the species tree**

* Started with a list of accession numbers that Rika made containing:
  + One representative genome per bacterial and archaeal order based on GTDB
  + Eukaryotes used by Chen et al. 2020 in Arsenic paper
  + Added 58 extra eukaryotic genomes from the Hug et al. 2016 paper
* Used GToTree to get prepare an alignment and preliminary tree that looked reasonable
  + Played with various flags and added more euks (the 58 from above) because for some reason, they seemed to fall off the alignments because they were too long.
  + The set of GToTree parameters that I used to make the final alignment was:

GToTree -a forGToTree\_evenmoreeuks.txt -H Universal\_Hug\_et\_al.hmm -G 0.5 -o 7.5GToTree\_output\_moreeuks -n 30 -j 30 -t -L

* + - -a is the input file of which is a list of accession numbers
    - -H gives the set of single copy genes (SCGs) that are pulled out. In this case the Hug set
    - -G filters out genomes that don’t have enough target genes. In this case, genomes are only included if they have ½ of the Hug SCGs
    - -O names the output file
    - -n tells GToTree how many CPUs to use. Using 30 CPUs makes this whole process run for 12-18 mins
    - -t and -L gets more helpful genome labels
  + GToTree creates a bunch of very helpful files located in the a new directory
    - The \*.tre file can be exported into IToL for visualization to evaluate topology
    - Aligned\_SCGs\_mod\_names.faa can be used for future tree construction
* Used RAxML to make a tree of the GToTree alignment

((Run RAxML in screen or you will be sorry))

raxmlHPC-PTHREADS-AVX -f a -m PROTGAMMALG -p 12345 -x 12345 -# 100 -s 7.5Alignment\_GotoTree.faa -n 7.5GtoTree\_eukarcout -T 30 -o [all the archaea and eukaryotes]

* + - Specifying the outgroup is a pain but necessary. I chose to set the archaea and eukaryotes as the outgroup instead of the bacteria since it’s fewer genomes
    - This tree took a few days to run

**Turning the Species Tree into a Chronogram**

* Created calibration files based on an extended discussion with Eva and Rika detailed in the appendix
* Converted the alignment from GToTree into a phylip format using the python script on the server:

convert\_afa\_to\_phy.py 7.5Aligned\_SCGs\_mod\_names.faa

* Ran phylobayes with the calibration files, phylip alignment, and tree file from RAxML, set root prior as LUCA

pb -d [phylip] -T[ RAxML\_bestTree] -cal [calibration.calib] -[model (ugam or ln)] [chain name] -rp [LUCA date] [deviation]

* + Ran with conservative and liberal calibration files each in log normal (-ln) and ugam modes.
  + For each permutation, ran two chains in parallel in different screen windows
* And then wait for a lifetime…
  + Phylobayes is not a super fast process. With a species tree of 871 species, expect to run for ~a month
  + You can check how converged your chains are using tracecomp (and in theory bpcomp, but that doesn’t work for me for some reason)

tracecomp -x [burn in] [interval of trees to sample] [name of run 1] [name of run 2]

* + - The guidelines in the phylobayes manual suggest that you want an effective size of >50 and a relative difference <0.3. Not sure if that is relevant to all parameters…
    - The arsenic paper uses a burn in of 20% of the cycles run
* Once the trees have converged (whatever that means) generate a chronogram

readdiv -x [burn in] 1 [chain name]

* I let the chains run >60,000 cycles and also plotted the trace files for loglik, length, and stat and estimated where they all seemed to level off— I ended up using 2500 which is also what Chris used.

**Using Annotree to Make a Gene Tree (For each gene)**

**Get the Species List from Annotree and Corresponding Make an Aligned Fasta**

* Go to [Annotree](http://annotree.uwaterloo.ca/) and click the big “Launch” button in the middle
* In the search bar up top enter the gene you are looking for. By default, you will be searching the KEGG database, though you can change that using the little drop down menu to the left of the search bar. Hit enter on your keyboard.
* Click “all query results” which will automatically download a csv file of all the gene hits. Rename this file to [Genename].csv and scp it onto the server.
* Do this for both the bacterial and archaeal trees! (You’ll have to change the drop-down menu at the top left.) Combine the results.
* Run annotree\_to\_curated\_fasta.py which will pull out just the genes from genomes in a provided alignment file \*\* for some reason, sometimes Annotree moves around the columns, so you may need to edit lines 51 and/or 57 in the script so that you index the “gtdbID” and “sequence” columns respectively. \*\*\*
* You will now have a file called [Genename].faa. Now align with MUSCLE:

muscle -in [Genename].faa -out [Genename].afa

* scp the new aligned file ([Genename].afa) back to your local computer

**Trim the Fasta with Phylemon**

* Go to the [Phylemon](http://phylemon2.bioinfo.cipf.es/) website and click on Utilities>Alignment Utilities>TrimAI
* Click browse server>Upload new file. Find your alignment. Select the data type as “Aligned sequences” and give your data a name. Click “Upload”.
* Under “Method” select “Automated 1” from the dropdown menu
* Give your job a name, leave default settings (don’t click the spurious sequences box), and hit “Run”
* Once the job is done (<1 min) it will turn green in the jobs list. Click on it
* Click on outfile.out and rename it something useful. I like [Genename].trimAI.afa.
* scp the new file back to baross

**Find the Best Evolutionary Model using IQ-Tree**

* Go to the [IQ-Tree](http://iqtree.cibiv.univie.ac.at/) website.
* Click on the “Model Selection” tab.
* Hit “Browse” and find your newly trimmed alignment.
* Leave the default settings and enter your email if you want to get updates on your job.
* Click “Submit Job”
* Once the job is done, click the “Full Result” tab and scroll down until you find “Best-fit model according to BIC.” This is your best model!

\*\*As of Dec 2020 the IQ-Tree web server does not work. You can download IQ-Tree onto your local box and then run iqtree -s [alignment file] -m TESTONLY\*\* Note: in response to reviewer requests, we modified some of the model-finding parameters, but this will depend largely on the sequences you are looking at.

**Turn the Fasta into a Phylip and Construct a Tree using RAxML -NG**

* Turn the alignment in afa form into a phy

convert\_afa\_to\_phy.py [Genename].trimAI.afa

* Make a tree in a screen window

raxml-ng --bootstrap --msa [Genename].trimAI.afa.phy --model [model selected by IQ-Tree] --prefix [Genename] --seed [random number]--threads [I used 4]

* + This tree will run to convergence or 1000 bootstraps, whichever occurs first
* If the tree runs to 1000 bootstraps, check for convergence (this may take a few minutes, so you may want to stay in screen)

raxml-ng --bsconverge --bs-trees [treename].raxml.bootstraps -- prefix [name to save the bscheck] --threads 1 --bs-cutoff 0.03

* If you have not converged, run another set of 1000 bootstraps the same way that you ran the first
* Concatenate the old set of bootstraps and the new set

cat [old set] [new set] > [combined file name]

* Check for convergence again...
* Rinse and repeat until you reach convergence!
* (note: for the final paper, in response to some reviewer requests, we switched to using IQ-TREE for both the model finding and the tree construction.)

**Using ecceTERA**

**Garrett Chappell\***

**Summer 2021**

\* text in pink added by Joanne Boden in 2022

\* text in blue added by Bryan Le in 2022

**Useful Links:**

* [ecceTERA Manual](https://github.com/celinescornavacca/ecceTERA/blob/master/ecceTERA.pdf)
* [ecceTERA GitHub](https://github.com/celinescornavacca/ecceTERA)
* [Paper introducing ecceTERA](https://academic.oup.com/bioinformatics/article/32/13/2056/1743358)
* [Paper introducing recPhyloXML](https://academic.oup.com/bioinformatics/article/34/21/3646/4995844)
* [recPhyloXML scripts GitHub](https://github.com/WandrilleD/recPhyloXML)
* [Website explaining recPhyloXML](http://phylariane.univ-lyon1.fr/recphyloxml/)

**Getting Ready to Run:**

* Test that ecceTERA is working using:
  + >> ecceTERA species.file=/usr/local/ecceTERA-master/tests/Stree.tree gene.file=/usr/local/ecceTERA-master/tests/HBG284008\_1.gtrees recPhyloXML.reconciliation=true
  + Should return:
    - Cost of a most parsimonious reconciliation: 85
* Input files required:
  + Chronogram (ex. Cons\_ugam.chronogram)
  + Bootstrapped gene tree (ex. aprA.raxml.bootstraps)

**Running ecceTERA:**

* In screen, run ecceTERA using:
  + >> ecceTERA species.file=./[chronogram file] gene.file=./[gene tree file] verbose=true print.reconciliations=1 recPhyloXML.reconciliation=true amalgamate=true
* What the different parameters mean:
  + verbose (optional): gives extra information while ecceTERA is running, sometimes it is nice to have this extra information to know what is happening
  + print.reconciliations: tells ecceTERA to make output files
    - =1: prints median reconciliations and a random reconciliation
    - =2: prints all reconciliations (many files)
  + recPhyloXML.reconciliation: makes the output files in recPhyloXML format, much better than the default ecceTERA output format
  + amalgamate: amalgamates all of the gene trees to compute the reconciliation instead of doing each of the bootstraps separately
  + There are also many other parameters you can change outlined in the manual, including being able to get the Newick trees ecceTERA used, changing the costs of different events, and changing the output file names/directories
  + Gene.file is the file containing the gene trees. If providing a sample of gene trees (like the bootstrapped trees from RaxML) then each tree should be on a separate line and in newick format. If any of the genes in the tree do not match up to a species in the chronogram file, then ecceTERA will not calculate the median reconciliations needed to determine when the gene originated.
* Immediately after beginning the ecceTERA run, detach from the screen and check top:
  + ecceTERA occasionally takes up huge amounts of RAM (%MEM in top) and it is important to watch top closely at the beginning to make sure it doesn’t start to use too much
  + %MEM will climb at about 0.3% for every top update until it stabilizes
    - It usually stabilizes and stops increasing within about 10 minutes
    - You can double-check that it stabilized by going back to the screen you started ecceTERA in
      * There should be a new line below the line “Computing ALE memory usage (MB): ###” that gives the matrix size, for example:
        + Computing ALE memory usage (MB): 262
        + 25815 x 380628 matrix
  + Once the %MEM stops increasing, it will stay at that amount of RAM usage for the whole run, probably only increasing a couple of tenths of a percent over the rest of the run
* The amount of time that the ecceTERA run will take could vary widely, most likely depending on the number of bootstraps in the gene tree
  + For example, one run with a species tree with 871 species and a gene tree with 1000 bootstraps took 7 hours
  + Another run with a species tree with 871 species and a gene tree with 2000 bootstraps took 1 day and 19 hours
  + Overall, they seem to take somewhere between 40 minutes and 2 days.

**Analyzing the Output:**

* There should be 3 output files from the run:
  + reconciliationsFile\_canonical\_asymmetric.recPhyloXML
  + reconciliationsFile\_canonical\_random.recPhyloXML
  + reconciliationsFile\_canonical\_symmetric.recPhyloXML
* The file I’ve used so far to do analysis is the “symmetric” recPhyloXML file
* Run the python script getInternalNodeDates.py:
  + Call the script using:
    - >> python getInternalNodeDates.py [chronogram file] [recPhyloXML file]

This command only works with Python2. This script is a standalone script, it can be copied into any directory to use

* + Takes the internal node names/numbers from the recPhyloXML ecceTERA output and matches them to the internal nodes of the chronogram to assign a range of dates for each node
  + Outputs a file with the name “[chronogram file]\_internal\_nodes.txt”
* Run the python script recPhyloXMLEventSummary.py:
  + - python ecceTERA/recPhyloXML/python/recPhyloXMLEventSummary.py -i [recPhyloXML file] -o symmetric.events --include.transfer.departure
  + Takes the recPhyloXML ecceTERA output file and returns a file that lists the number of each type of event at each node
  + Outputs a file with the name “symmetric.events”
* Open the “symmetric.events” file in BBEdit using “Open from FTP/SFTP Server”
  + Use Ctrl/Cmd + F to find and replace all the double spaces with single spaces until there are no double spaces left (only a single space between columns)
  + Replace all of the single spaces with tabs (\t)
  + Save and close
  + Do this automatically on the command line using this command: sed -r -i "s|\s+|\t|g" symmetric.events
* Run the python script recphyloxmlinterpreter.py:
  + Call script using:
    - >> python **recphyloxmlinterpreter.py** [symmetric.events file] [chronogram internal nodes txt file] [recPhyloXML file]
  + Takes the list of events that occurred, the dates for each internal node, and the recPhyloXML file and returns a file that gives dates for each event
  + If you get an error on line 72 saying “list index out of range”, it’s because the symmetric.events file hasn’t been reformatted properly to leave a single tab between each column
  + Outputs a file with the name “symmetric.events\_event\_dates.txt”
* You can use the python script getGeneBirthData.py to get info on the oldest event:
  + Call the script using:
    - >> python getGeneBirthDate.py [symmetric.events\_event\_dates.txt file]
  + It will print data about the oldest event (or one of the oldest events if there are multiple tied for the oldest)
    - Assumes the oldest event identified by ecceTERA is near the birth of the gene
* The “symmetric.events\_event\_dates.txt” file can now be analyzed or visualized using R or python
  + I used R to make histograms using:
    - data <- read.delim("[symmetric.events\_event\_dates.txt filepath]", na.strings="?")
    - ggplot(data, aes(x=midpoint.date)) + geom\_histogram(aes(y=(..count../[total number of events]), fill=event), binwidth = 250, boundary = 0) + xlim(4000,0) + labs(x="Million years ago", y="Proportion of total events", title="[gene] Gene Events") + scale\_fill\_manual(values=c("#ffd166", "#06d6a0", "#118ab2"), name="Type of Event", labels=c("Duplication", "Horizontal Gene Transfer", "Loss")) + theme(legend.position = "bottom", plot.title = element\_text(size = rel(1.75), hjust = 0.5), axis.line.y = element\_line(linetype = 1)) + geom\_hline(yintercept = 0)
    - ggsave("[gene]\_histogram.jpg")

**Troubleshooting:**

* ecceTERA uses too much RAM:
  + Try reducing the number of bootstraps in the gene tree (might need to reduce by a lot)
* ecceTERA lists tons of genes “not taken into account”:
  + Check the raxml.bootstraps file for “-” and replace all of them with “..” using BBEdit
* ecceTERA returns error “Segmentation fault (core dumped)”:
  + If there are many genes listed as “not taken into account,” replace the “-” in the raxml.bootstraps file with “..” and that could fix it

**Appendix: Sulfur Tree Calibration Events**

|  |  |  |
| --- | --- | --- |
| Calibration Events | Conservative (Ga) | Liberal (Ga) |
| LUCA (set as root prior)\* | >3.8 (1) | >3.8 (1) |
| Origin of Methanogenesis\* | >2.7 (3) | >3.51 (4) |
| Origin of Oxygenic Photosynthesis\* | >2.45 (7) | >3 (8) |
| Origin of Eukaryotes\* | >1.7 (9) | >3.2 (10) |
| Origin of Plastids/Rhodophytes diverge\* | >1.05 (11) | >1.2 (12) |
| Akinetes diverge from cyanobacteria lacking cell differentiation | >1.0 (13) | >1.5 (14) |
| Origin of Animals | >0.635 (15) | >0.635 (15) |

\*Calibration events included in nitrogen analysis

**References**

1. S. J. Mojzsis, *et al.*, Evidence for life on Earth before 3,800 million years ago. *Nature* **384**, 55–59 (1996).

2. E. A. Bell, P. Boehnke, T. M. Harrison, W. L. Mao, Potentially biogenic carbon preserved in a 4.1 billion-year-old zircon. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14518–14521 (2015).

3. J. L. Eigenbrode, K. H. Freeman, Late Archean rise of aerobic microbial ecosystems. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 15759–15764 (2006).

4. J. M. Wolfe, G. P. Fournier, Horizontal gene transfer constrains the timing of methanogen evolution. *Nat. Ecol. Evol.* **2**, 897–903 (2018).

5. M. R. Warke, *et al.*, The Great Oxidation Event preceded a Paleoproterozoic “snowball Earth.” *Proc. Natl. Acad. Sci.* **117**, 202003090 (2020).

6. H. C. Betts, *et al.*, Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).

7. A. Bekker, *et al.*, Dating the rise of atmospheric oxygen. *Nature* **427**, 117–120 (2004).

8. T. W. Lyons, C. T. Reinhard, N. J. Planavsky, The rise of oxygen in Earth’s early ocean and atmosphere. *Nature* **506**, 307–315 (2014).

9. K. Pang, *et al.*, The nature and origin of nucleus-like intracellular inclusions in Paleoproterozoic eukaryote microfossils. *Geobiology* **11**, 499–510 (2013).

10. E. J. Javaux, C. P. Marshall, A. Bekker, Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. *Nature* **463**, 934–938 (2010).

11. T. M. Gibson, *et al.*, Precise age of Bangiomorpha pubescens dates the origin of eukaryotic photosynthesis. *Geology* **46**, 135–138 (2018).

12. N. J. Butterfield, Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386–404 (2000).

13. K. Pang, *et al.*, Nitrogen-Fixing Heterocystous Cyanobacteria in the Tonian Period. *Curr. Biol.* **28**, 616-622.e1 (2018).

14. S. Golubic, V. N. Sergeev, A. H. Knoll, Mesoproterozoic Archaeoellipsoidès: akinetes of heterocystous cyanobacteria. *Lethaia* **28**, 285–298 (1995).

15. G. D. Love, *et al.*, Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature* **457**, 718–721 (2009).