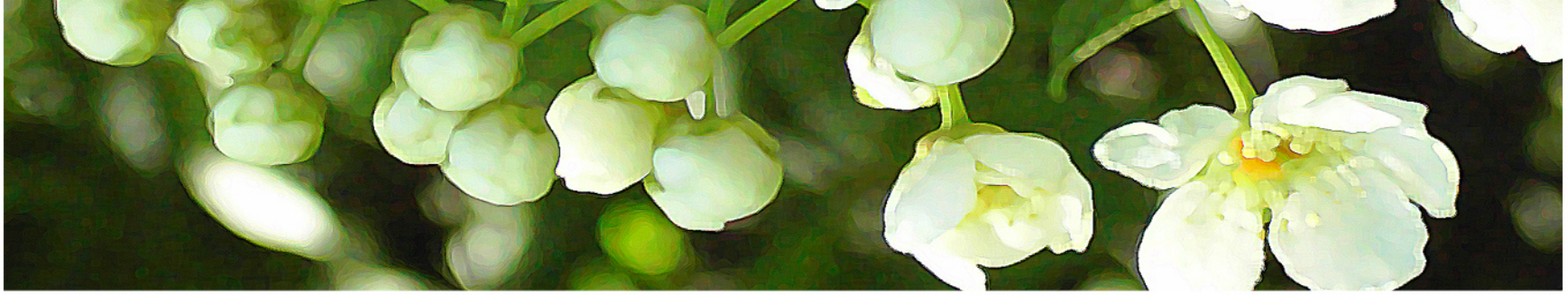


UiO : Department of Biosciences
University of Oslo

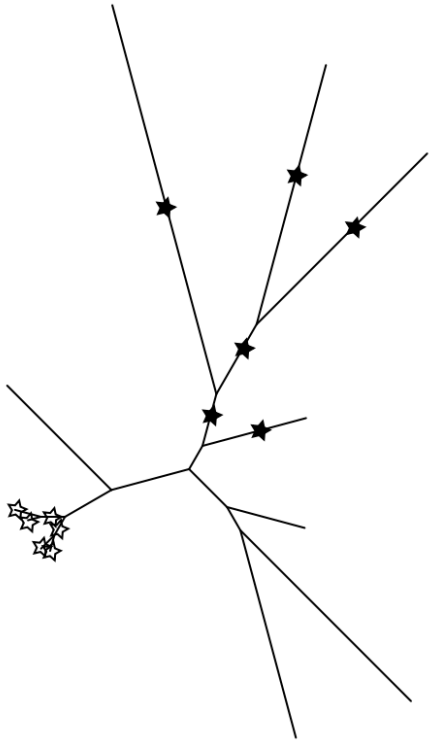
Phylogeny - Evolutionary placement algorithm

Anders K. Krabberød

a.k.krabberod@ibv.uio.no

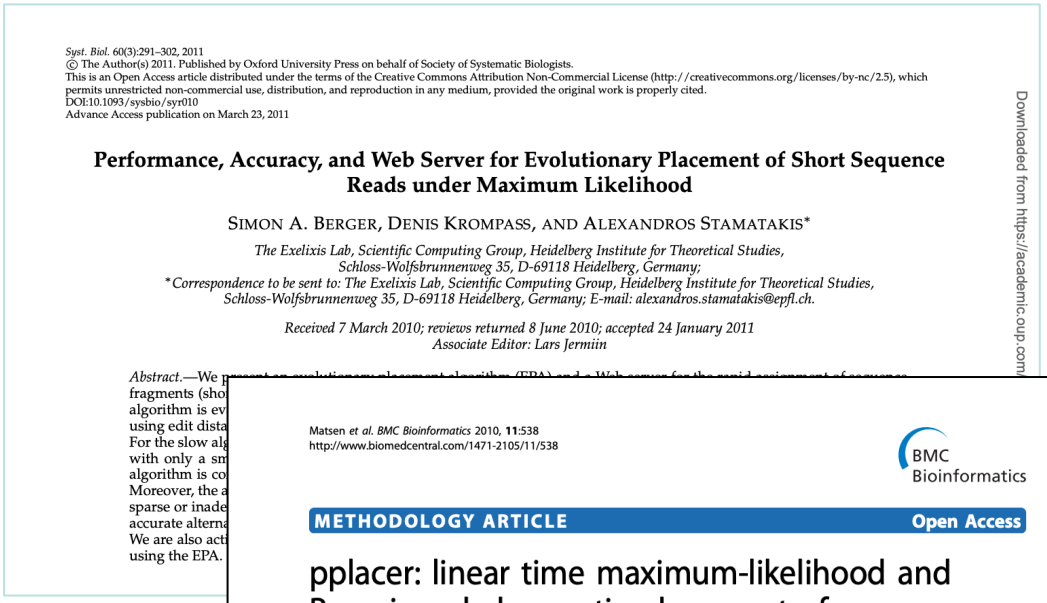
Evolutionary Placement Algorithm

- Phylogenetically identification of short query sequences (illumina, 454 etc) using a set of full length reference sequences and a reference tree.
- Developed as an alternative to taxonomic assignment based on sequence identity (blast) or homology (hmmer)



EPA - Evolutionary Placement Algorithm

- EPA developed by the Exelixis lab run by A. Stamatakis; Berger et al. (2011)
- Similar to *pplacer* published *by* Matsen et al. (2010)



Syst. Biol. 60(3):291–302, 2011

DOI:10.1093/sysbio/syr010
Advance Access publication on March 23, 2011

Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood

SIMON A. BERGER, DENIS KROMPASS, AND ALEXANDROS STAMATAKIS*

The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany;
**Correspondence to be sent to: The Exelixis Lab, Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, D-69118 Heidelberg, Germany; E-mail: alexandros.stamatakis@epfl.ch.*

Received 7 March 2010; reviews returned 8 June 2010; accepted 24 January 2011
Associate Editor: Lars Jermiin

Abstract.—We p... fragments (sho... algorithm is ev... using edit dista... For the slow alg... with only a sm... algorithm is co... Moreover, the a... sparse or inade... accurate alterna... We are also acti... using the EPA.

Matsen *et al. BMC Bioinformatics* 2010, **11**:538
http://www.biomedcentral.com/1471-2105/11/538

BMC Bioinformatics

METHODOLOGY ARTICLE **Open Access**

pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree

Frederick A Matsen[1*], Robin B Kodner[2,3], E Virginia Armbrust[2]

Abstract

Background: Likelihood-based phylogenetic inference is generally considered to be the most reliable classification method for unknown sequences. However, traditional likelihood-based phylogenetic methods cannot be applied to large volumes of short reads from next-generation sequencing due to computational complexity issues and lack of phylogenetic signal. "Phylogenetic placement," where a reference tree is fixed and the unknown query sequences are placed onto the tree via a reference alignment, is a way to bring the inferential power offered by likelihood-based approaches to large data sets.

Results: This paper introduces pplacer, a software package for phylogenetic placement and subsequent

EPA - Evolutionary Placement Algorithm

- From the Exelixis lab run by A. Stamatakis; Berger (2011)
- Similar to *pplacer* (Matsen 2010)
- Jointly they defined the output format of Phylogenetic Placements in 2012
-

A small example

```
{
  "tree": "((A:0.2{0},B:0.09{1}):0.7{2},C:0.5{3}){4};",
  "placements":
  [
    {"p":
      [[1, −2578.16, 0.777385, 0.004132, 0.0006],
      [0, −2580.15, 0.107065, 0.000009, 0.0153]
      ],
    "n": ["fragment1", "fragment2"]
    },
    {"p": [[2, −2576.46, 1.0, 0.003555, 0.000006]],
      "nm": [["fragment3", 1.5], ["fragment4", 2]]}
  ],
  "metadata":
  {"invocation":
    "pplacer -c tiny.refpkg frags.fasta"
  },
  "version": 3,
  "fields":
  ["edge_num", "likelihood", "like_weight_ratio",
        "distal_length", "pendant_length"]
}
```

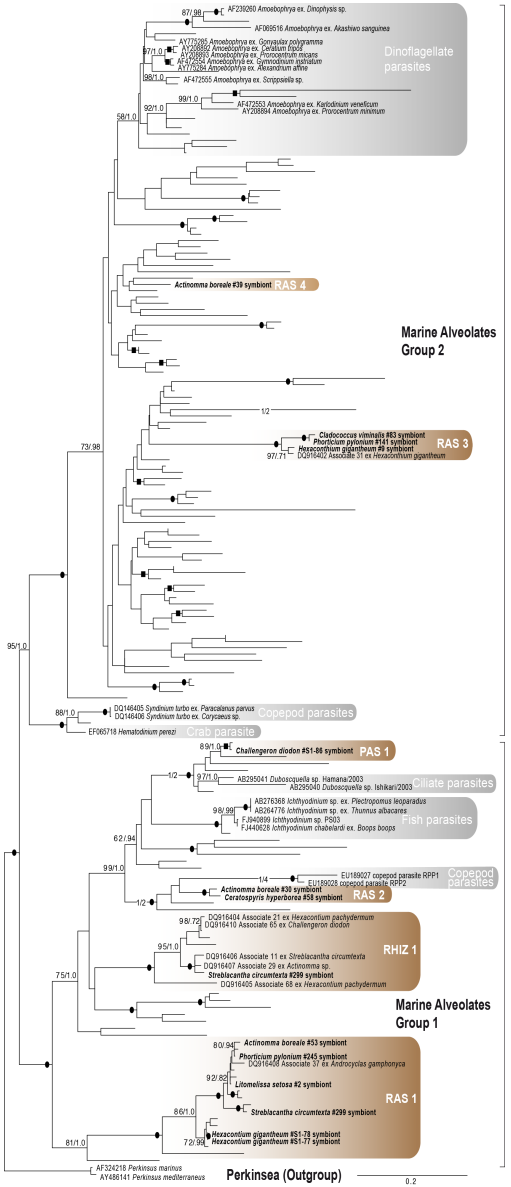
Matsen et al. (2012)

Why EPA?

- Taxonomic assignments with BLAST have shortcomings. Especially when the reference database lacks matches for the query sequence.
  - Uncertainties in blast hits are hard to evaluate.
  - What does a low identity score mean?
  - What does a low e-value mean?
    - Dependent on fragment length and database size.
  - Is the hit from the database correctly annotated?

Why EPA?

- Phylogeny can be used as an alternative and complementary method to do taxonomic assignment.
- Sequences are placed together based on the evolutionary relationship, not annotations from a database.
- (if an evolutionary model is used as basis for the phylogenetical tree)



Bråte et al (2012). ML tree of Marine Alveolates, 18S.

Why EPA?

- Phylogenies from many short sequences are hard to do because:

Why EPA?

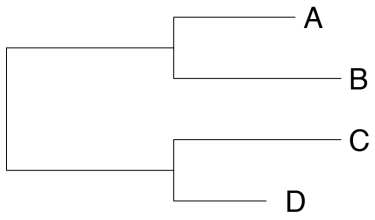
- Phylogenies from many short sequences are hard to do because:
  - Short sequences in general have low phylogenetic resolution.
  - Often the short fragment used in amplicon sequencing has particularly low phylogenetical signal since it is chosen because of its variability.

Why EPA?

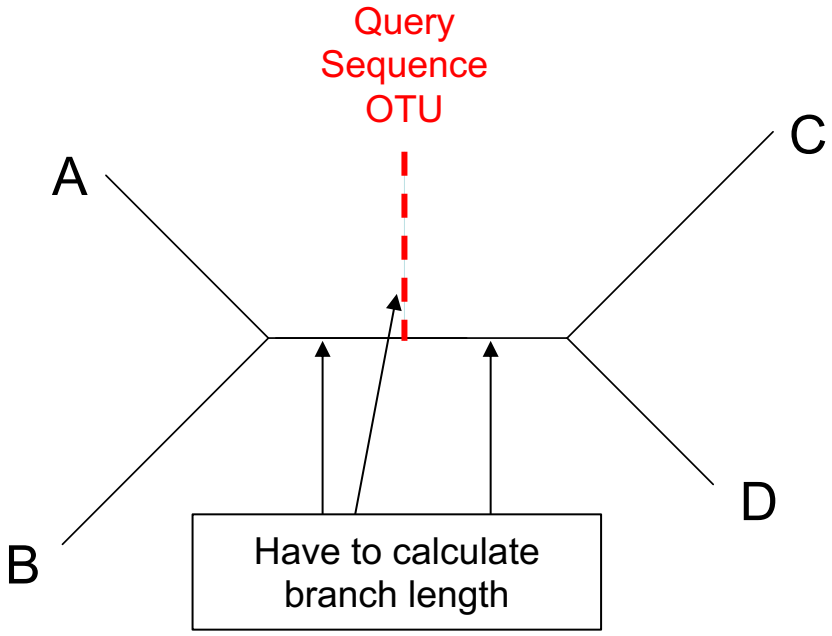
- Phylogenies from many short sequences are hard to do because:
  - The possible rearrangements of a fully resolved phylogenetic tree grows exponentially with the number of taxa added.
  - The number of unrooted trees for *n* taxa: $(2n-5)!/[2^{n-3}*(n-3)!]$
  - Number of rooted trees for *n* taxa: $(2n-3)!/[2n-2*(n-2)!]$

Number of Taxa	Number of unrooted trees	Number of rooted trees
3	1	3
5	15	105
10	$2.02*10^{6}$	$3.45*10^{7}$
20	$2.22*10^{20}$	$8.20*10^{21}$
80	$2.18*10^{137}$	$3.43*10^{139}$

How EPA Works

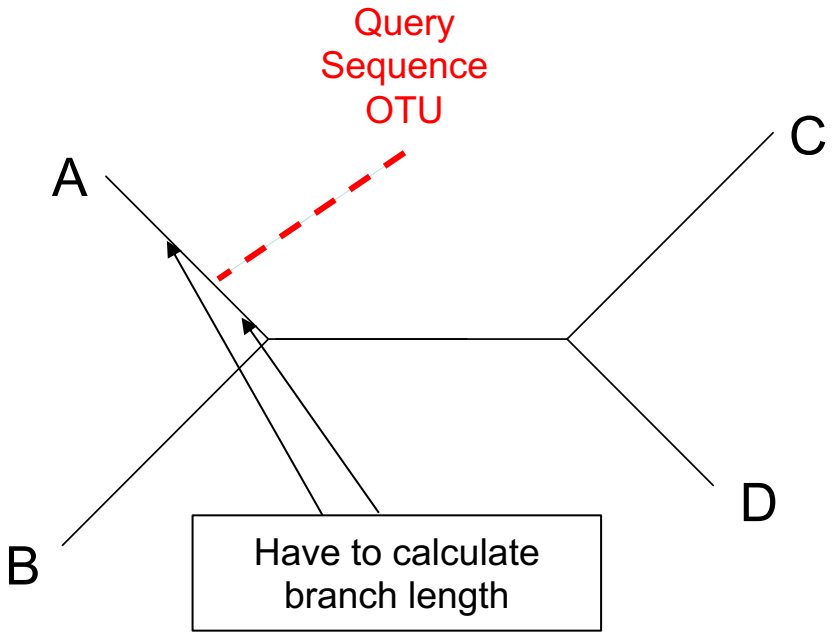


- Query sequences are placed on a reference tree individually by a searching algorithm that tries to find the branch with the best likelihood for the query one at the time.



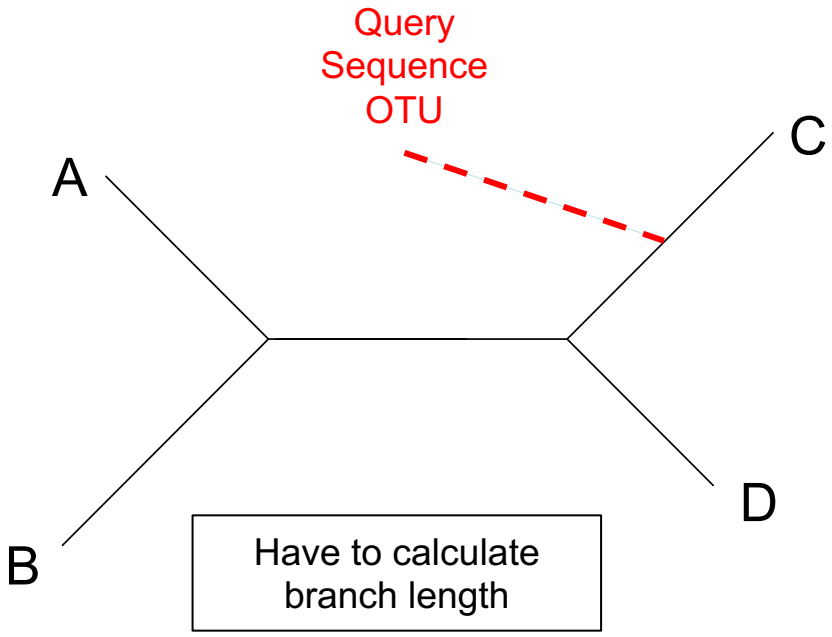
How EPA Works

- Query sequences are placed on a reference tree individually by a searching algorithm that tries to find the branch with the best likelihood for the query one at the time.



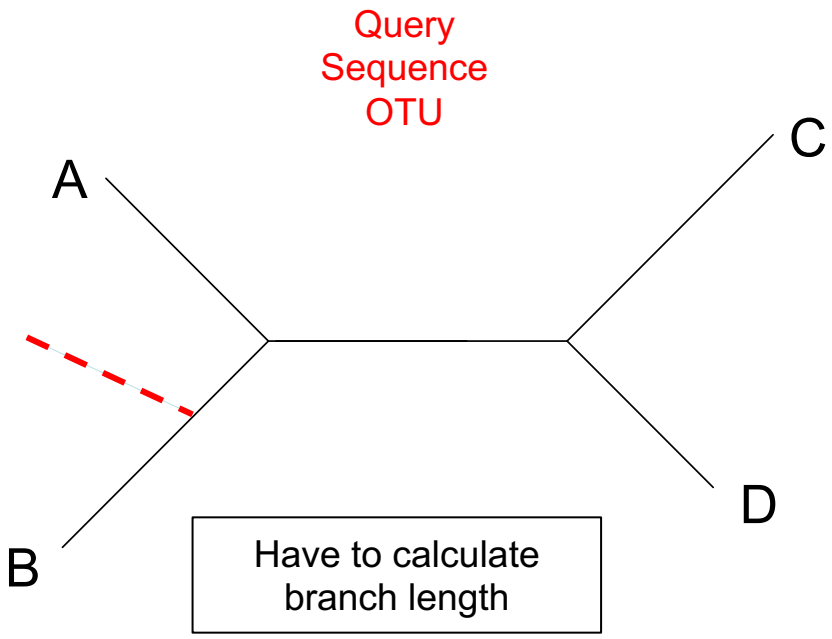
How EPA Works

- Query sequences are placed on a reference tree individually by a searching algorithm that tries to find the branch with the best likelihood for the query one at the time.



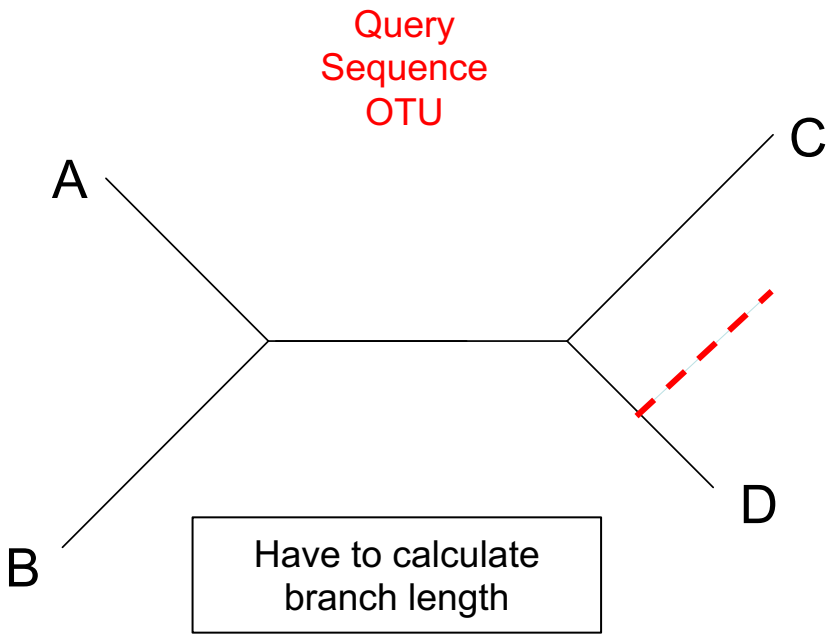
How EPA Works

- Query sequences are placed on a reference tree individually by a searching algorithm that tries to find the branch with the best likelihood for the query one at the time.



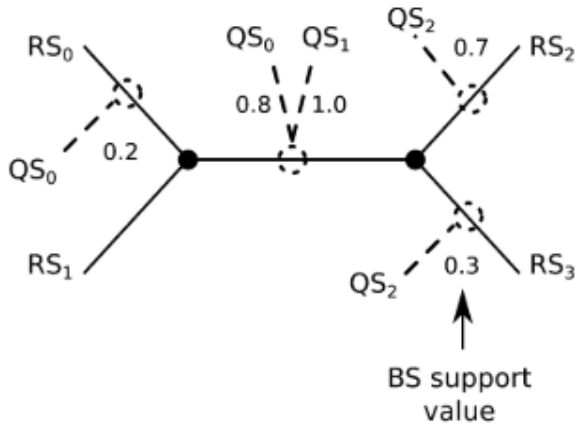
How EPA Works

- Query sequences are placed on a reference tree individually by a searching algorithm that tries to find the branch with the best likelihood for the query one at the time.
- Since this is done individually for each sequence it is easy to do parallelisation

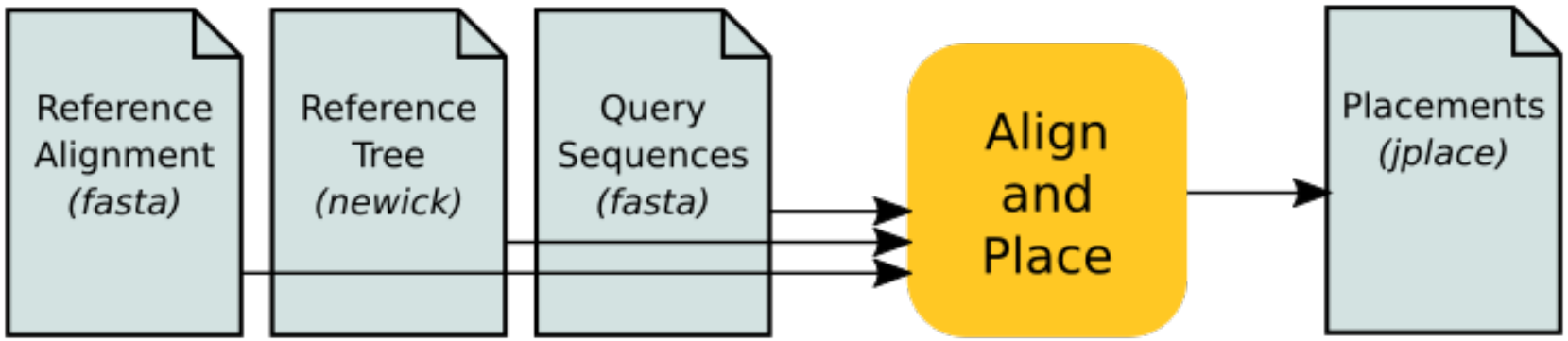


How EPA Works

- One query sequence (OTU) can have several placemen
- The likelihood for all placements as well as the likelihood weight ratio (lwr) is reported for each OTU.



Workflow



Sequencens from GenBank, Pr2 etc. Aligned with mafft, muscle

RAxML, IQ-tree, Phylobayes etc.

Output from DADA2, usearch, swarm, Illumina data without clustering

- Short read alignment (mafft. PaPaRa)
- EPA or pplacer

Post processing with Gappa (or Guppy, Genesis and others)

https://github.com/lczech/gappa/wiki/Phylogenetic-Placement

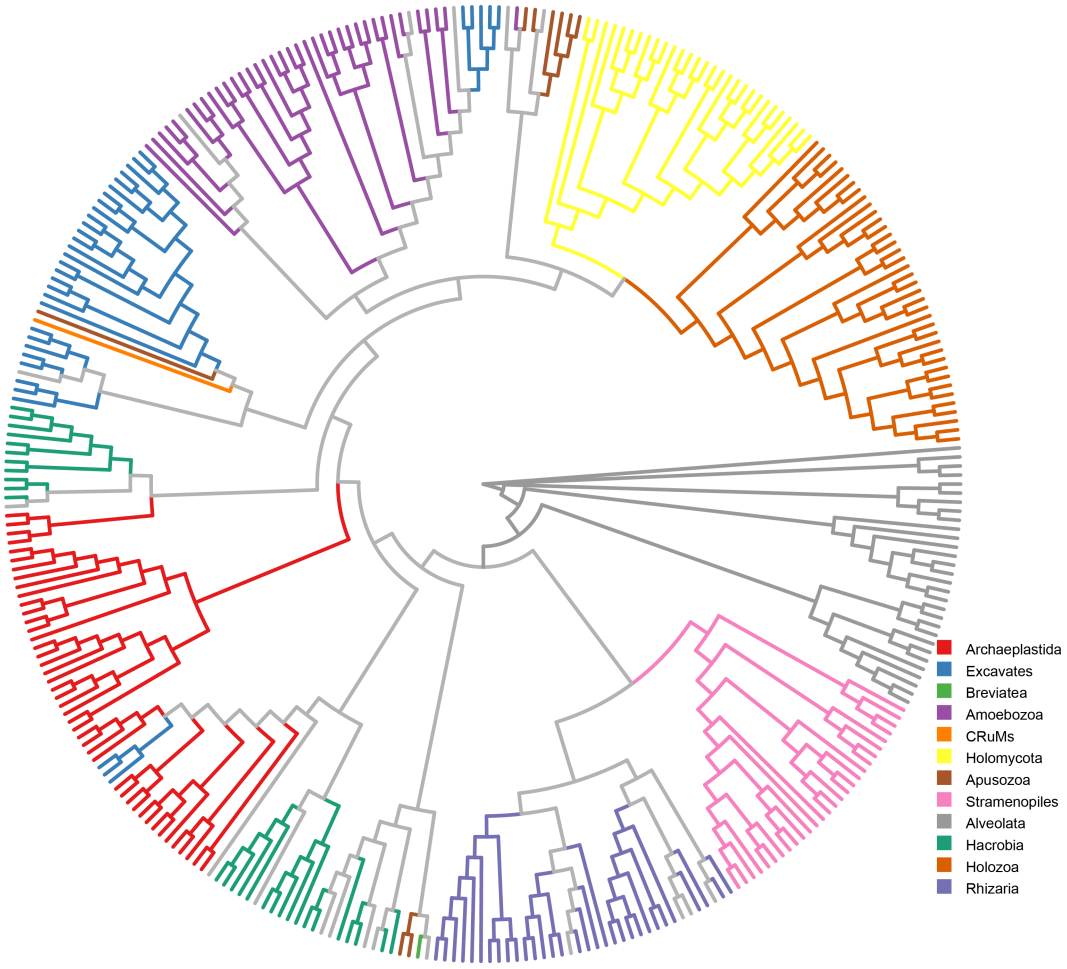
Workflow

- Alignment with reference sequences for the reference tree
  - Several programs: MAFFT (good for 18S in nt), Muscle (good for coding genes in aa)
- Reference tree.
  - A tree where you want to place your own OTUs.
  - Can be made to suit different needs.
  - For instance a global tree with representatives from all major groups
  - A tree specific for a phylogenetic group of interest
- The model used to infer the reference tree
- Alignment with the added OTUs
  - Mafft --addfragment, PaPaRa

Workflow – Reference tree

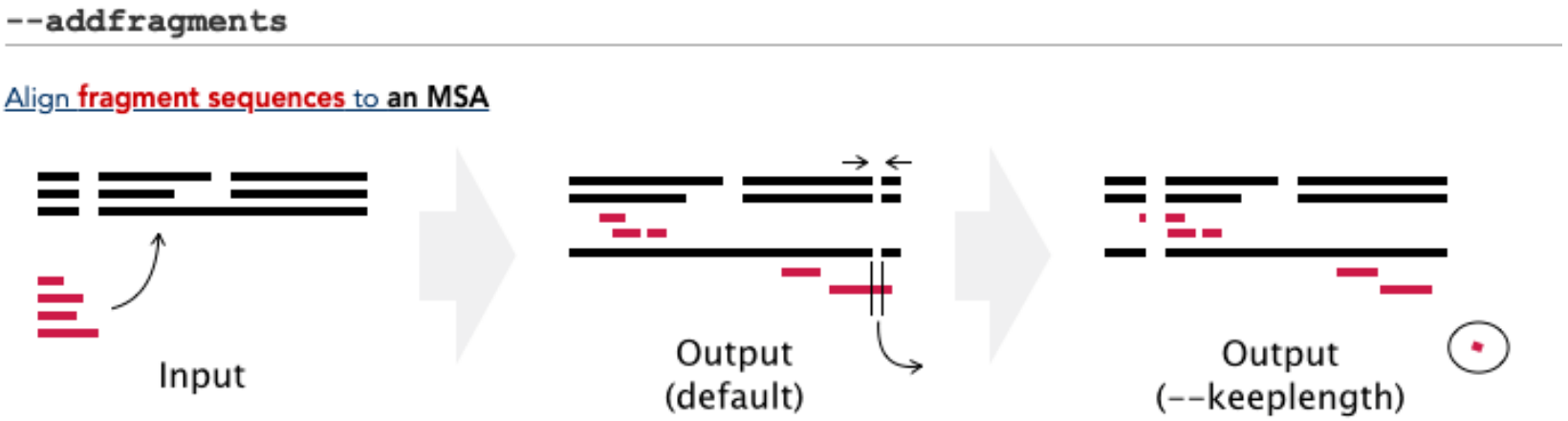
- Reference tree
- Should consist of taxa relevant for your study
  - Depends on the question you want answered
- Example :
  - Alignment with mafft (linsi algorithm)
  - Global phylogenetic tree with 331 taxa from the major eukaryotic groups, retrieved from GenBank
  - Maximum likelihood tree made with RAxML, GTRGAMMA model and 300 bootstraps

Workflow – Reference tree



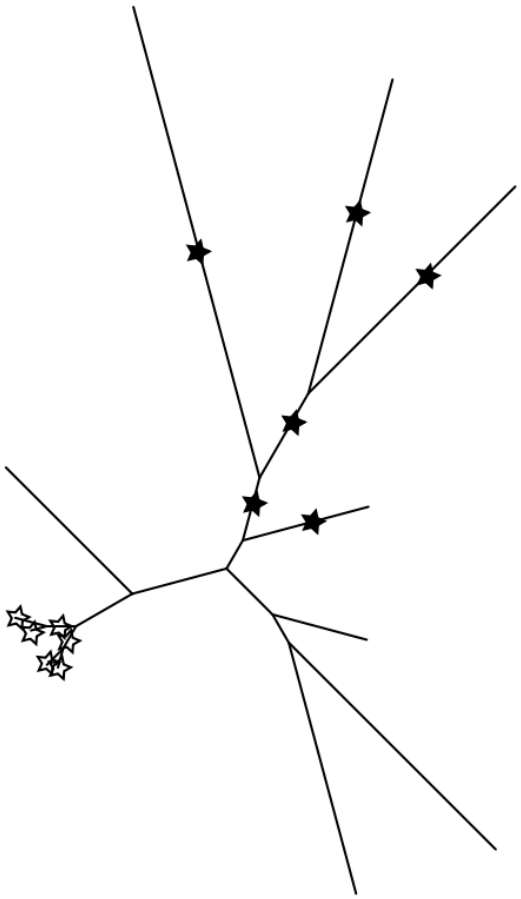
Workflow – Add short sequences to Alignment

- Add the shorter OTUs
  - mafft --addfragment OTUs input > output



Workflow – Place OTUs

- RAxML reads the full alignment and the reference tree.
- The names of the reference sequences in the alignment and the tips in the tree have to be identical.
- RAxML then places the sequences on the reference tree and calculates the likelihood of the placement.



Output : jplace file

A small example


```
{
   "tree": "((A:0.2{0},B:0.09{1}):0.7{2},C:0.5{3}){4};",
   "placements":
   [
      {"p":
         [[1, −2578.16, 0.777385, 0.004132, 0.0006],
         [0, −2580.15, 0.107065, 0.000009, 0.0153]
         ],
      "n": ["fragment1", "fragment2"]
      },
      {"p": [[2, −2576.46, 1.0, 0.003555, 0.000006]],
         "nm": [["fragment3", 1.5], ["fragment4", 2]]}
   ],
   "metadata":
   {"invocation":
      "pplacer -c tiny.refpkg frags.fasta"
   },
   "version": 3,
   "fields":
   ["edge_num", "likelihood", "like_weight_ratio",
            "distal_length", "pendant_length"]
}
```


The reference tree

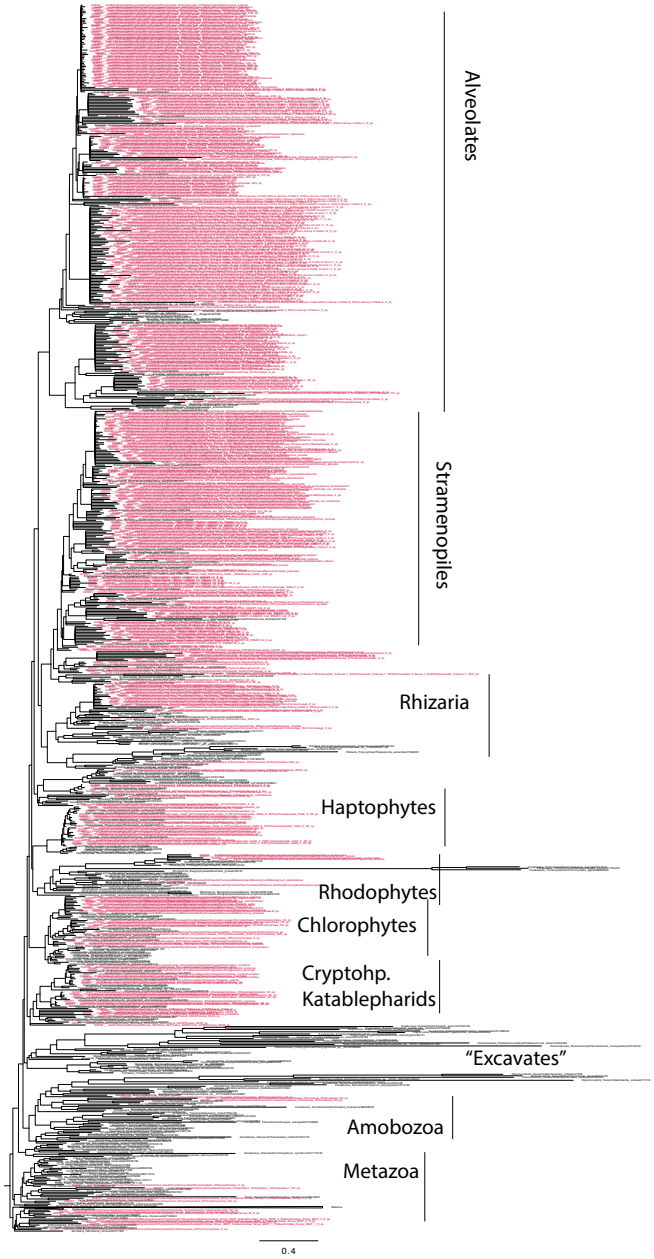
Placement information for each query sequence. Each new entry starts with “p”

Optional metadata: in RAxML the command is printed

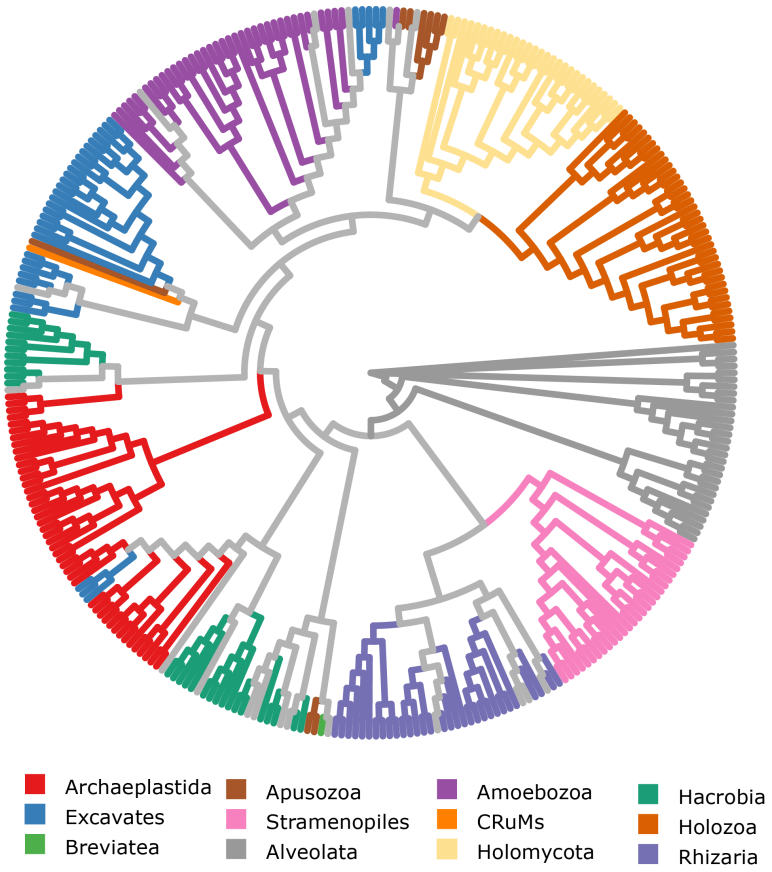
Version and field definitions

Full tree

- 313 reference sequences in black.
- Placed sequences in red



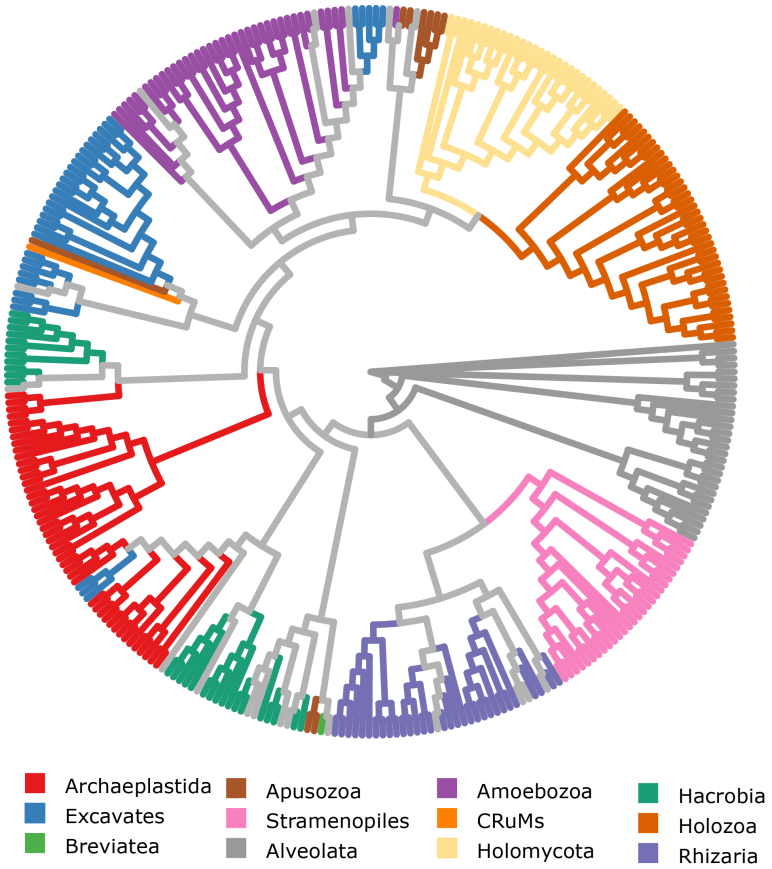
Reference tree



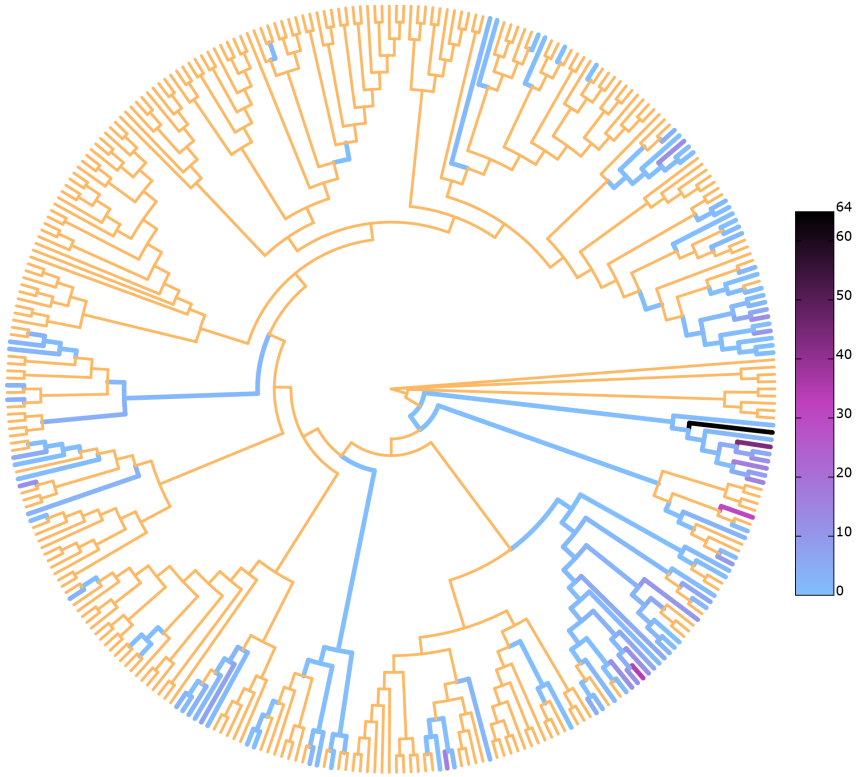
- 313 full length 18S sequences covering the major branches of the tree of life.
- Tree inferred with GTRGAMMA model in RAxML

UiO : Department of Biosciences
University of Oslo

Reference tree



OsloFjord OTUs



Gappa,
https://github.com/lczech/gappa/wiki

Example

- F. Mahé et al. (2017)
- 10 567 804 amplicons (V4)
- Reference tree with 512 full length 18S sequences
- RAxML backbone tree
- PaPaRa alignment

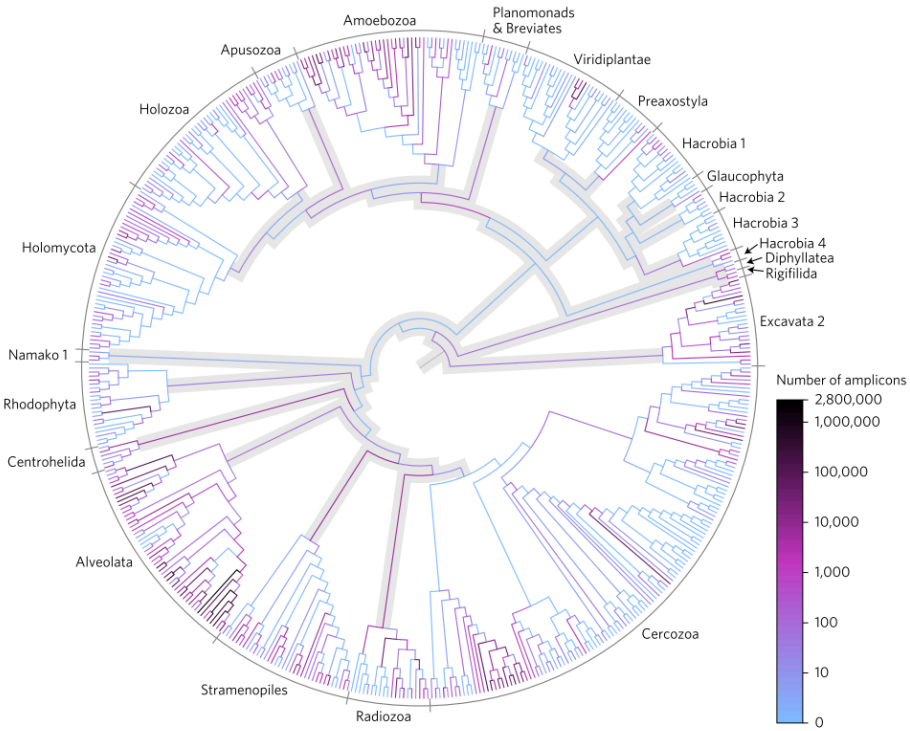


Figure 2 | Phylogenetic placement of Neotropical soil protist reads on a taxonomically unconstrained global eukaryotic tree. Reads were dereplicated into strictly identical amplicons. Inferred relationships between these major taxa may differ from those obtained with phylogenomic data. Alveolata includes Apicomplexa and Ciliophora; Holozoa includes animals; Holomycota includes fungi. Branches and nodes outside of known clades are shaded grey. In our conservative approach only OTUs that placed within known clades with high likelihood-weight scores were retained.