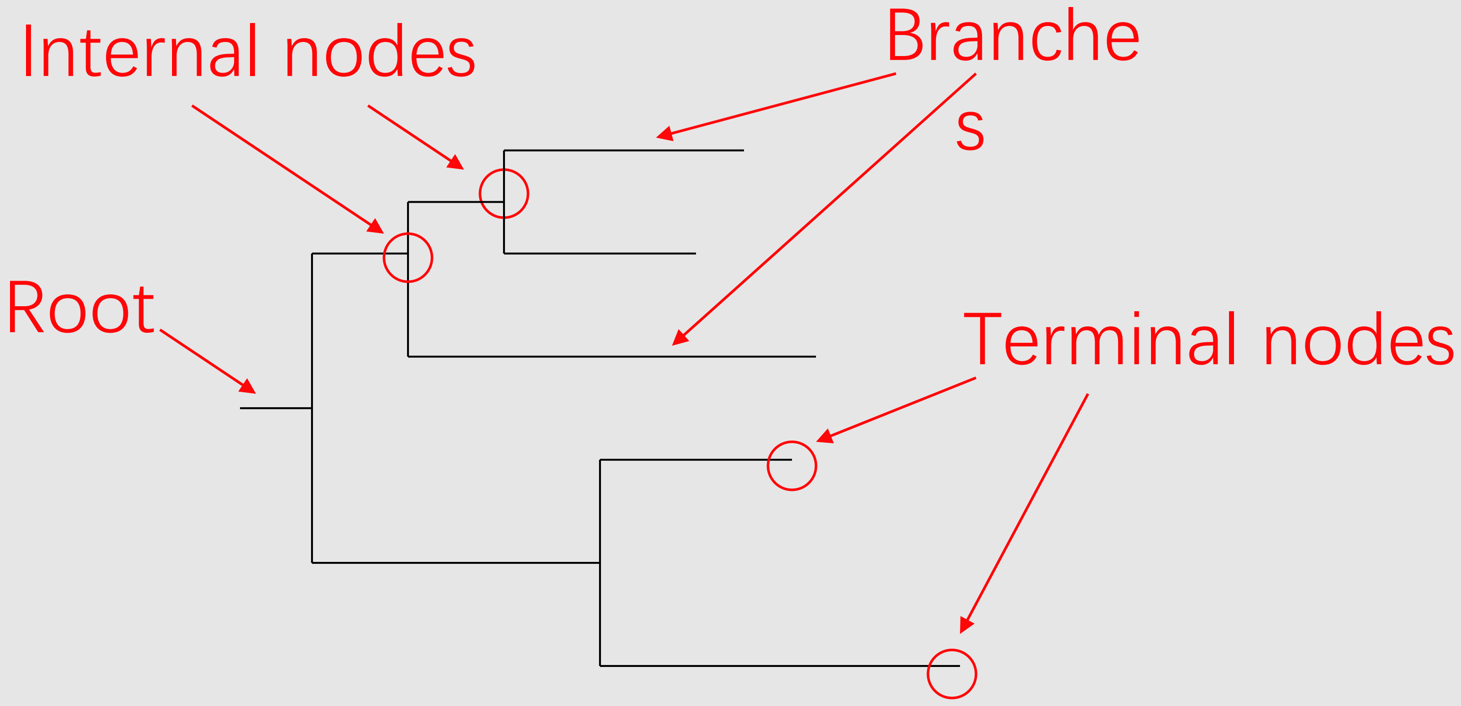
2019workshop session1-tree reconstruction

Reporter: Guoxing Yin

Tree (in biology)

- Dendrogram representing the evolutionary relationship between species.

Terminology



Phylogram or phylogenetic tree

What kind of trees

- Gene trees
- Species trees
- Time trees
- Etc.

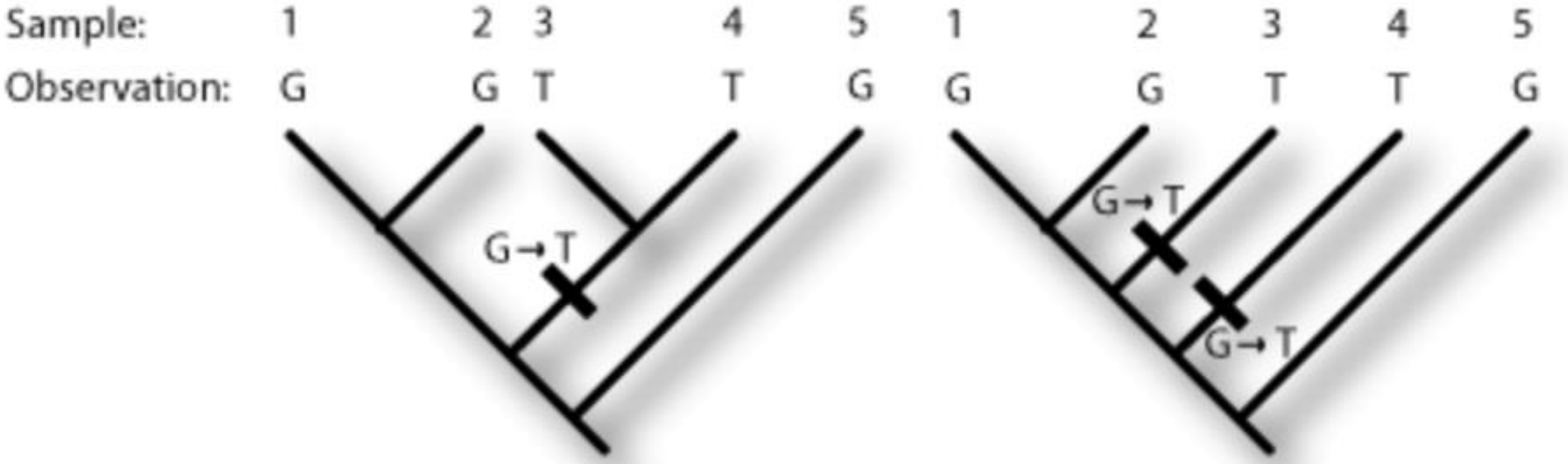
Based on the molecular level of phylogenetic methods can be divided into two categories

- Discrete feature based approach
- Example：Maximum parsimony methods, Maximum likelihood methods, Bayesian methods, etc .
- Based on distance methods
- Example :N-J methods, etc.

Maximum parsimony methods (MP)

- maximum parsimony is an optimality criterion under which the phylogenetic tree that minimizes the total number of character-state changes is to be preferred.
- Under the maximum-parsimony criterion, the optimal tree will minimize the amount of homoplasy .
- In other words, under this criterion, the shortest possible tree that explains the data is considered best.
- MP is not consistent, particularly in the case of unequal evolutionary rates between different lineages.

Using Maximum Parsimony to Choose Between Two Possible Trees



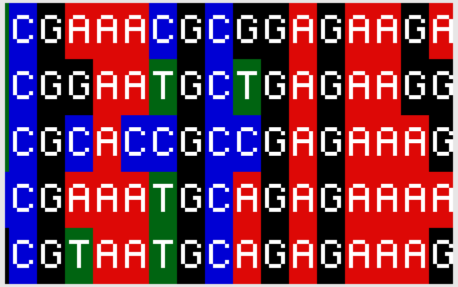
1 change required
→ better tree

2 changes required
→ poorer tree

Maximum likelihood methods(ML)

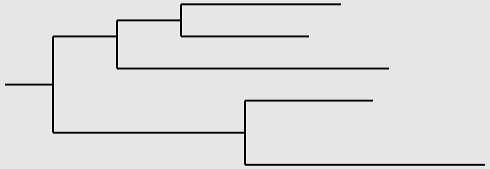
- A completely statistical-based phylogenetic tree reconstruction approach that takes into account the probability of each nucleotide substitution in each set of comparisons.
- The tree with the largest sum of probabilities is most likely the most real phylogenetic tree.
- The important advantage of probabilistic methods over parsimony is statistically consistent.

Basic evolutionary models



```
Taxa 1 CGAAACGCGGAGAAGA
Taxa 2 CGGAATGCTGAGAAGG
Taxa 3 CGCACCGCCGAGAAAG
Taxa 4 CGAAATGCAGAGAAAA
Taxa 5 CGTAATGCAGAGAAAG
```

- Topology and branch length



- Substitution matrix

r_{TC} (= r_{CT}), r_{TA} (= r_{AT}), r_{TG} (= r_{GT})
r_{CA} (= r_{AC}), r_{CG} (= r_{GC})
r_{AG} (= r_{GA})

- Stationary base frequencies

f_T, f_C, f_A, f_G,

$$L_D = \Pr(D \mid H)$$

Bayesian methods

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P\left(B_j\right)P\left(A|B_j\right)}$$

Example

- There have an example, suppose there a lot of people in gym, 40% girls and 60% boys, all boys wear pants, half girls wear pants and half girls wear skirt if you see someone wear pants randomly, what is the probability of a girl?
- P(A)= probability of someone wear pants=80%
- P(B)= probability of someone is girl=40%
- P(AIB)= probability of girl wearing pants=50%
- P(BIA)= P(B)P(AIB)/P(A)=25%

Distance methods

- we calculate the distance matrix between every two sequences, repeat the merging of the two sequences with the shortest distance, and finally construct the optimal tree.

Gene trees

- "Gene" trees represent the evolutionary history of the genes included in the study.
- Gene trees can provide evidence for gene duplication events, as well as speciation events.
- Sequences from different homologs can be included in a gene tree; the subsequent analyses should cluster orthologs, thus demonstrating the evolutionary history of the orthologs.

Concatenated tree

- Concatenated tree use concatenate independence gene, so the tree are more truly reflect the evolutionary history of species.
- We use raxml to reconstruct concatenated tree.

RAxML (Randomized Axelerated Maximum Likelihood)

- It is a popular program for phylogenetic analysis of large datasets under maximum likelihood.
- Its major strength is a fast maximum likelihood tree search algorithm that returns trees with good likelihood scores.

Usage

-
```
raxmlHPC-PTHREADS  –T=12 –p=12345 –m=GTRGAMMA –s=***.phy –n=raxml -y -f a –x 12345  -# 100 –q=partation
```

- -T number of nodes
- -p parsimony Random Seed
- -m substitution Model
- -s sequence File Name
- -n output File Name
- **-**f a rapid Bootstrap analysis and search for best-scoring ML tree in one program run
- –x Specify an integer number (random seed) and turn on rapid bootstrapping
- -# Specify the number of alternative runs on distinct starting trees

-model

- GTRCAT: General Time Reversible + Optimization of substitution rates + Optimization of site specific evolutionary rates.

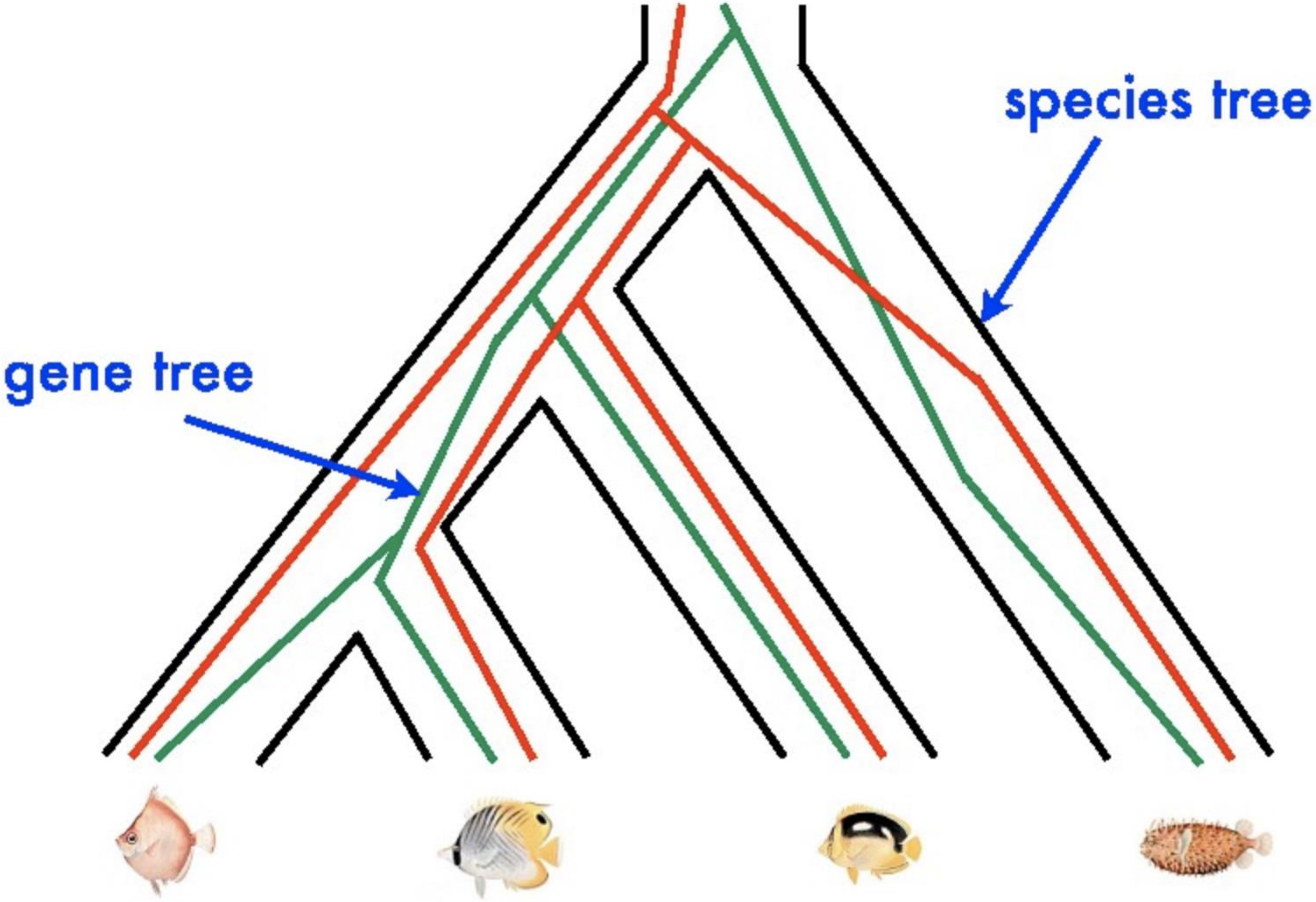
- GTRGAMMA: : General Time Reversible + Optimization of substitution rates + GAMMA model of rate heterogeneity

ExaML

- Exascale Maximum Likelihood (ExaML) code for phylogenetic inference on supercomputers using MPI.
- This code implements the popular RAxML search algorithm for maximum likelihood based inference of phylogenetic trees.
- It uses a radically new MPI parallelization approach that yields improved parallel efficiency, in particular on partitioned multi-gene or whole-genome datasets.

Species trees

- "Species" trees recover the genealogy of taxa, individuals of a population, etc.
- Internal nodes represent speciation or other taxonomic events.
- Species trees should contain sequences from only orthologous genes.


species tree
gene tree

ASTRAL

- ASTRAL is a tool for estimating an unrooted species tree given a set of unrooted gene trees.
- ASTRAL is statistically consistent under the multi-species coalescent model.
- ASTRAL finds the species tree that has the maximum number of shared induced quartet trees with the set of gene trees, subject to the constraint that the set of bipartitions in the species tree comes from a predefined set of bipartitions.

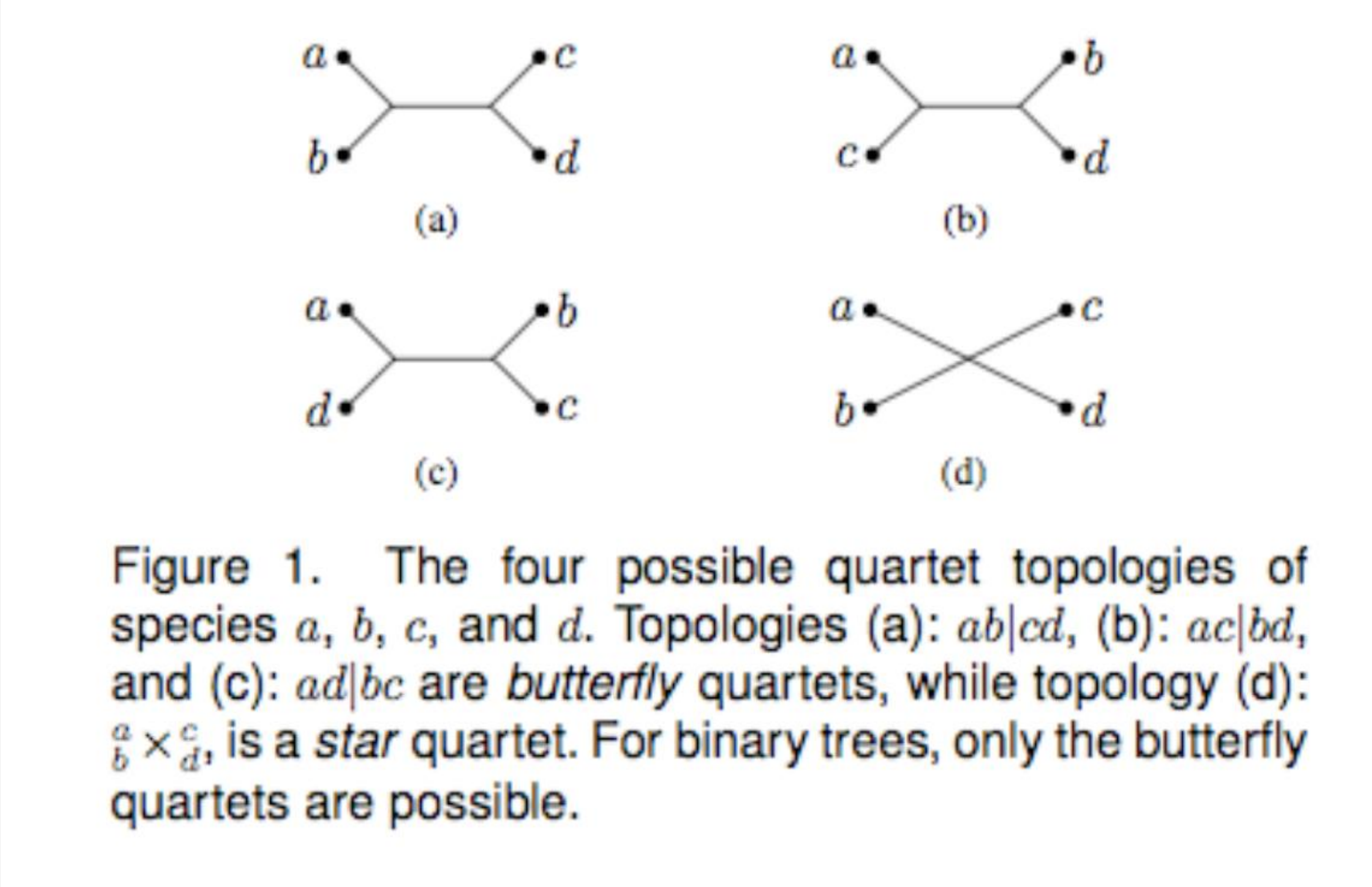


Figure 1. The four possible quartet topologies of species a, b, c, and d. Topologies (a): $ab|cd$, (b): $ac|bd$, and (c): $ad|bc$ are *butterfly* quartets, while topology (d): $\genfrac{}{}{0pt}{}{a}{b}\times\genfrac{}{}{0pt}{}{c}{d}$, is a *star* quartet. For binary trees, only the butterfly quartets are possible.

Usage

- Java –jar astral.jar –i catree.tre –a speciesname.txt –o ***.tre –t 2

- -jar version of astral
- -i input file
- -a species name
- -o output file
- -t 2 full annotation

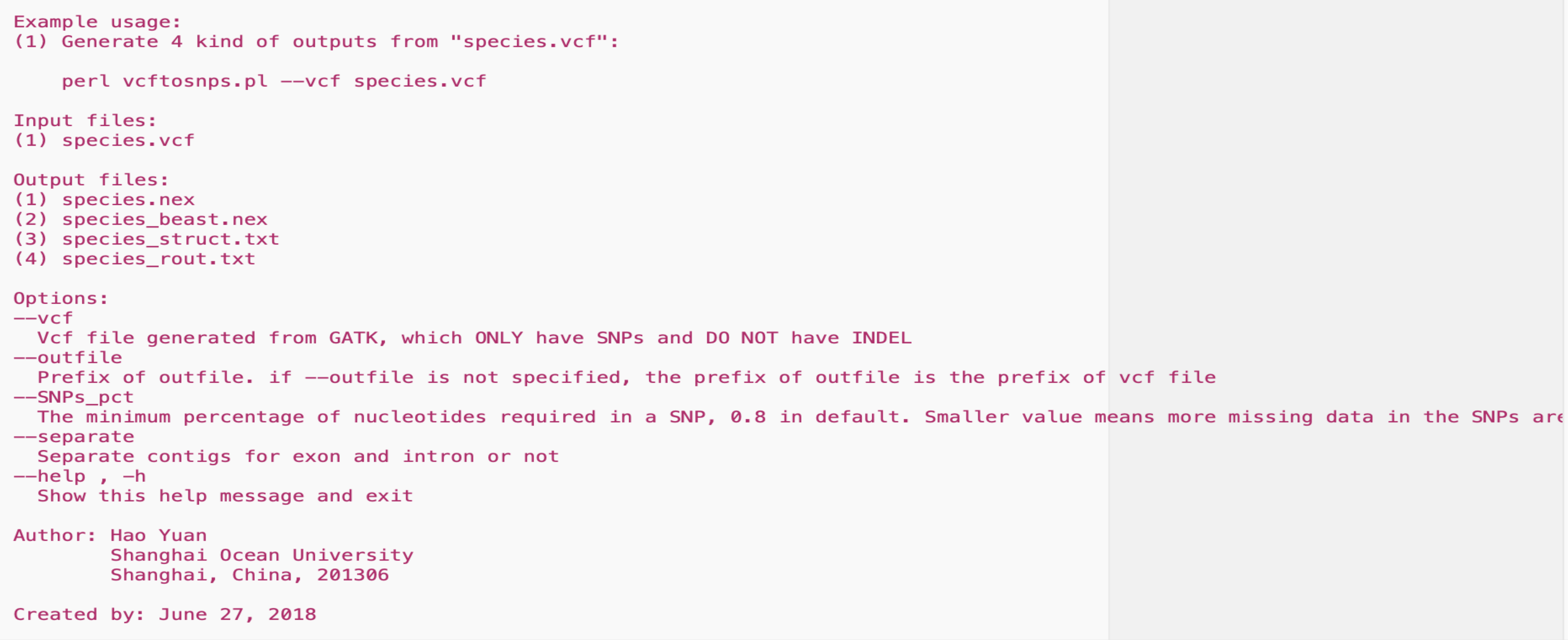
Beast

- BEAST is a cross-platform program for Bayesian inference using MCMC of molecular sequences. It is entirely orientated towards rooted, time-measured phylogenies inferred using strict or relaxed molecular clock models.
- BEAST uses MCMC to average over tree space, so that each tree is weighted proportional to its posterior probability.

SNAPP

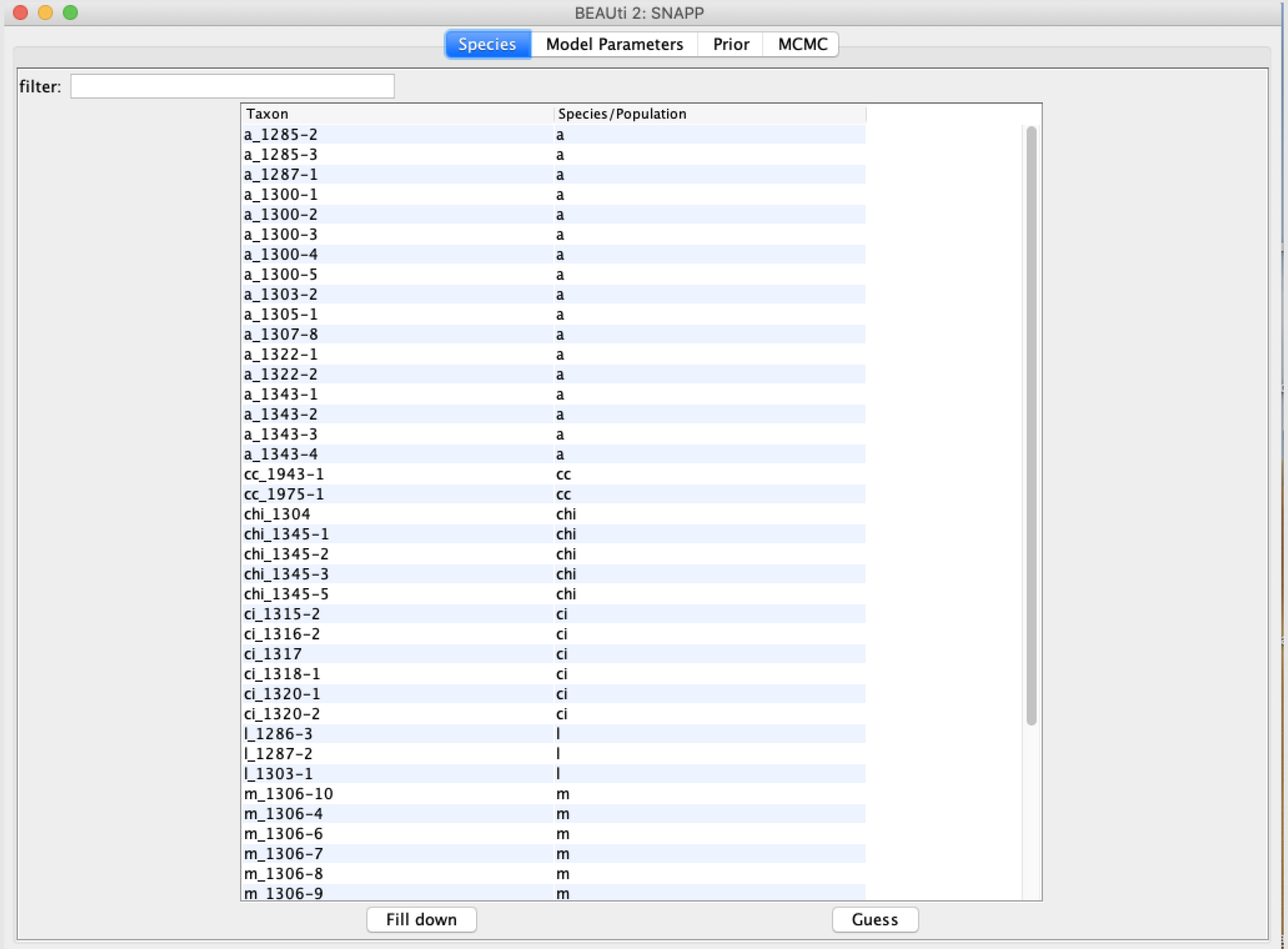
- SNAPP (SNP and AFLP Package for Phylogenetic analysis) is package for inferring species trees and species demographics from independent (unlinked) biallelic markers such as well spaced SNPs.
- It implements a full coalescent model, but uses a novel algorithm to integrate over all possible gene trees, rather than sampling them explicitly.

Usage



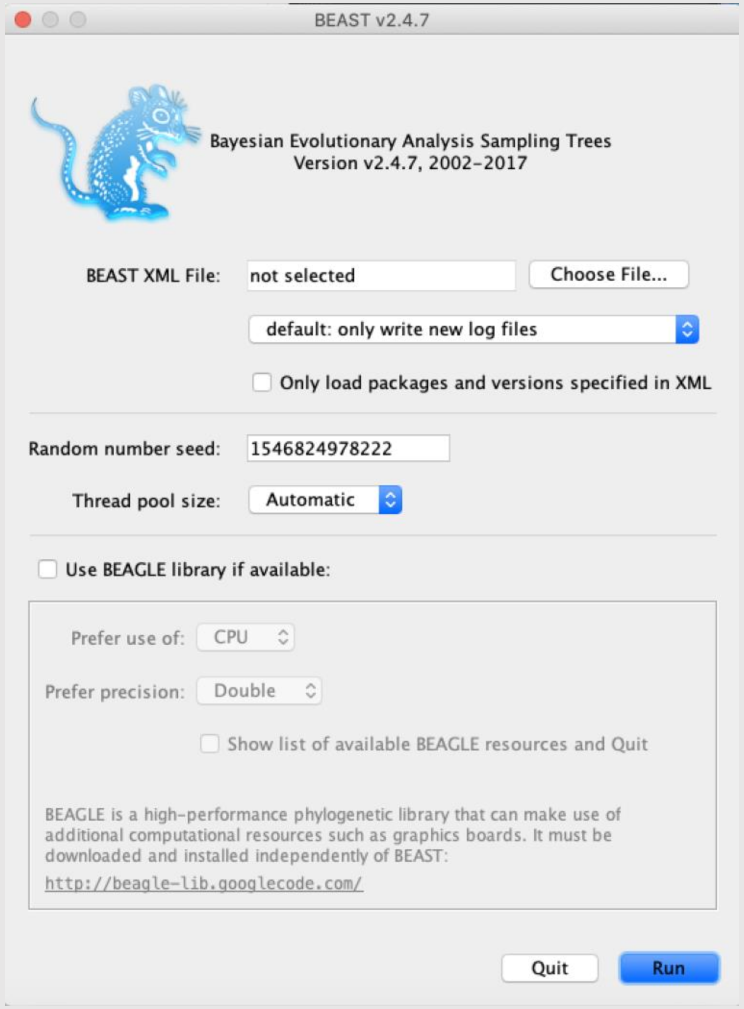
```
Example usage:
(1) Generate 4 kind of outputs from "species.vcf":

    perl vcftosnps.pl --vcf species.vcf

Input files:
(1) species.vcf

Output files:
(1) species.nex
(2) species_beast.nex
(3) species_struct.txt
(4) species_rout.txt

Options:
--vcf
  Vcf file generated from GATK, which ONLY have SNPs and DO NOT have INDEL
--outfile
  Prefix of outfile. if --outfile is not specified, the prefix of outfile is the prefix of vcf file
--SNPs_pct
  The minimum percentage of nucleotides required in a SNP, 0.8 in default. Smaller value means more missing data in the SNPs are
--separate
  Separate contigs for exon and intron or not
--help , -h
  Show this help message and exit

Author: Hao Yuan
        Shanghai Ocean University
        Shanghai, China, 201306

Created by: June 27, 2018
```

Setting parameter in snapp



Taxon	Species/Population
a_1285-2	a
a_1285-3	a
a_1287-1	a
a_1300-1	a
a_1300-2	a
a_1300-3	a
a_1300-4	a
a_1300-5	a
a_1303-2	a
a_1305-1	a
a_1307-8	a
a_1322-1	a
a_1322-2	a
a_1343-1	a
a_1343-2	a
a_1343-3	a
a_1343-4	a
cc_1943-1	cc
cc_1975-1	cc
chi_1304	chi
chi_1345-1	chi
chi_1345-2	chi
chi_1345-3	chi
chi_1345-5	chi
ci_1315-2	ci
ci_1316-2	ci
ci_1317	ci
ci_1318-1	ci
ci_1320-1	ci
ci_1320-2	ci
l_1286-3	l
l_1287-2	l
l_1303-1	l
m_1306-10	m
m_1306-4	m
m_1306-6	m
m_1306-7	m
m_1306-8	m
m_1306-9	m

Run in beast



BEAST v2.4.7

Bayesian Evolutionary Analysis Sampling Trees
Version v2.4.7, 2002-2017

BEAST XML File: not selected

default: only write new log files

□ Only load packages and versions specified in XML

Random number seed: 1546824978222

Thread pool size: Automatic

□ Use BEAGLE library if available:

Prefer use of: CPU

Prefer precision: Double

□ Show list of available BEAGLE resources and Quit

BEAGLE is a high-performance phylogenetic library that can make use of additional computational resources such as graphics boards. It must be downloaded and installed independently of BEAST:
http://beagle-lib.googlecode.com/

Thanks for your attention!