Polymorphism-aware phylogenetic models

Workshop, MIC-Phy 2021

Dominik Schrempf

February 16, 2021

Eötvös Loránd
University

```python
import numpy as np
import pandas as pd
import scipy as sp

import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sb

sb.set()
sb.set_style("ticks", {"axes.grid": True})
sb.set_context("notebook")
%config InlineBackend.figure_format = 'svg'
```

Introduction

This workshop is available on GitHub.

Our goal is to understand how we can use polymorphism-aware phylogenetic models (PoMo) to improve inferences from population data.

In the course of this workshop, we will infer a phylogenetic tree from test data using IQ-TREE2.

If you need help, please interrupt me anytime!

Preparation - Command line shell

Basic knowledge of the command line shell of your choice is assumed.

- If you do not know basic commands such as `cd`, `ls`, or `less`, just lean back and listen to the presentation.
- Otherwise, try to follow the steps and complete the workshop yourself.

In case you are lost:

- Have a look at the manual pages, if they exist.

```
man less
```

- Read how commands are used.

```
less --help
```

Preparation - Download workshop and data

Option 1: If you have `git` installed, use it.

```
git clone https://github.com/pomo-dev/micphy-workshop.git
cd micphy-workshop
```

Option 2: Manually download the archive (requires `wget`, and `unzip`).

```
wget https://github.com/pomo-dev/micphy-workshop/archive/master.zip
unzip master.zip
cd micphy-workshop-master
```

The advantage of Option 1 is that you can:

- update your working tree if I have to change something during the workshop; use `git pull`;
- reset to the initial state if you mess up; use `git reset --hard HEAD` (be careful, this erases all changes made by you).

Preparation - Install IQ-TREE2

Option 1: Install from the repository of your distribution. For example, use the Arch Linux User Repository.

```
yay -S iqtree
aura -A iqtree
```

Option 2: Compile yourself (not shown).

Option 3: Use `nix-shell` and the `shell.nix` expression provided in the base directory of the repository (requires `nix`).

```
nix-shell
```

```
Welcome to the MIC-Phy PoMo workshop.
The following version of IQ-TREE2 is available:
IQ-TREE multicore version 2.1.2 COVID-edition for Linux 64-bit built Jan  1 1980
Developed by Bui Quang Minh, James Barbetti, Nguyen Lam Tung,
Olga Chernomor, Heiko Schmidt, Dominik Schrempf, Michael Woodhams.
```

Preparation - Install IQ-TREE2

Option 4: Download the binary executable from the IQ-TREE2 homepage.

```
wget https://github.com/iqtree/iqtree2/releases/download/v2.1.2/iqtree-2.1.2-Linux.tar.gz
```

Make sure that you have permission to execute the file (`chmod +x`), and that the executable is in your `PATH` (or that you provide the path during execution).

```
tar -xzvf iqtree-2.1.2-Linux.tar.gz
chmod +x iqtree-2.1.2-Linux/bin/iqtree2  # Should not be necessary, but who knows.
mv iqtree-2.1.2-Linux/bin/iqtree2 ~/bin/ # If ~/bin is in your PATH.
```

Exercise - Test IQ-TREE2 version

IQ-TREE2 version

Check that your IQ-TREE2 version agrees with the one I am using.

```
iqtree2 --version
# /path/to/iqtree2 --version
# ./relative/path/to/iqtree2 --version
```

```
IQ-TREE multicore version 2.1.2 COVID-edition for Linux 64-bit built Jan  1 1980
Developed by Bui Quang Minh, James Barbetti, Nguyen Lam Tung,
Olga Chernomor, Heiko Schmidt, Dominik Schrempf, Michael Woodhams.
```

Exercise - Access IQ-TREE2 help

IQ-TREE2 help

Access the IQ-TREE2 help, maybe read through some command line flags.

```
iqtree2 --help | less
```

```
IQ-TREE multicore version 2.1.2 COVID-edition for Linux 64-bit built Jan  1 1980
Developed by Bui Quang Minh, James Barbetti, Nguyen Lam Tung,
Olga Chernomor, Heiko Schmidt, Dominik Schrempf, Michael Woodhams.

Usage: iqtree [-s ALIGNMENT] [-p PARTITION] [-m MODEL] [-t TREE] ...

GENERAL OPTIONS:
  -h, --help           Print (more) help usages
```

Fruit fly data

We are going to analyze some fruit fly data. The data comprises nine *Drosophila* populations obtained from PopFly[1].

NTH	Netherlands
EG	Egypt
FR	France
GA	Gabon
GU	Guinea
EF	Ethiopia
KN	Kenyia
SB	South Africa (Barkly East)
SP	South Africa (Phalaborwa)

[1] Hervas et al. (2017); thanks Rui for providing the counts files.

Exercise - Explore fruit fly data

Explore data

Have a look at the data in the `./data` folder.

```
data_description.csv
fruit_flies_10000.cf
fruit_flies_10000.consensus.fasta
fruit_flies_10000.random.fasta
fruit_flies_1000.cf
fruit_flies_1000.consensus.fasta
fruit_flies_1000.random.fasta
```

We have data of two different lengths (1k and 10k sites), and also in counts file and FASTA file formats.

Exercise - Run DNA substitution model

Before running PoMo, we will use a normal DNA substitution model.

DNA substitution model

- Infer a phylogenetic tree using a DNA substitution model.
- Explore the output files. Specifically have a look at the `.log`, the `.iqtree` and the `.treefile` files.

```
iqtree2 -nt 4 -redo -mredo -s fruit_flies_10000.consensus.fasta -B 1000
```

Questions

- Which substitution model was used? How was it determined?
- What is the determined maximum log likelihood?
- How does the tree look like (topology, bootstrap values, branch lengths)?

Exercise - Run PoMo I

PoMo

- Infer a phylogenetic tree using PoMo.
- Explore the output files.

```
iqtree2 -nt 4 -redo -s fruit_flies_1000.cf -m HKY+F+P -B 1000
```

Questions

- What is the average number of samples per population? What is the estimated heterozygosity? *Why is it important to check the heterozygosity?*
- Which virtual population size was used?
- What is the determined maximum log likelihood?
- How does the tree look like (topology, bootstrap values, branch lengths)?

Exercise - Run PoMo II

PoMo parameters

- Play around with different virtual population sizes.
- Use gamma rate heterogeneity.
- Compare different DNA substitution models.

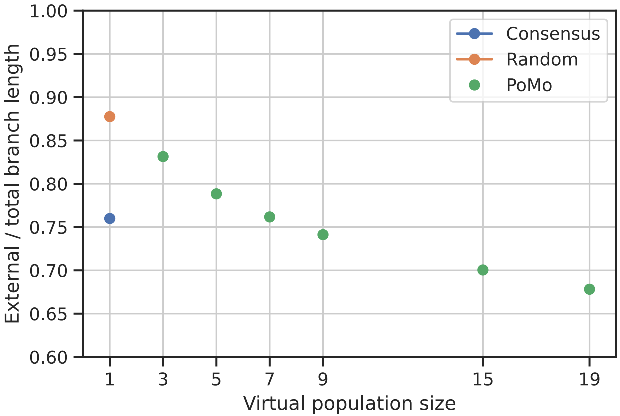
```
f=fruit_flies_1000.cf; m="HKY+F+P+N09+G2"; iqtree2 -nt 4 -redo -s $f -m $m -pre $f.$m
```

Questions

- What are the absolute and relative differences between branch lengths for different virtual population sizes?
- How do your results compare to using normal DNA substitution models. For a fair comparison, you have to run PoMo on the proper data with 10k sites.

Results

Results are provided in the `./results` folder.



Advice

Sometimes, the inference is unsuccessful. This may have several reasons. Two of them are:

- The likelihood derivative is zero or close to zero and numerical underflow occurs. This is especially an issue when N is large. Try using `-safe` (which is slower).
- The algorithm diverges. Try repeating the analysis with a different seed.

In general, it is recommended to perform replicate analyses and compare the parameters and log likelihoods. Further, a good starting tree can save a lot of time and worries.

Literature

PoMo	De Maio et al. (2015).
Reversible PoMo	Schrempf et al. (2016).
Non-reversible PoMo	Schrempf and Hobolth (2017).
Advanced models with PoMo	Schrempf et al. (2019).
IQ-TREE2	Minh et al. (2020).
Consistency of PoMo	Borges and Kosiol (2020).
PoMo with selection	Borges et al. (2019).

Bibliography I

- Borges, Rui, Gergely J. Szöllősi, and Carolin Kosiol (2019). “Quantifying GC-Biased Gene Conversion in Great Ape Genomes Using Polymorphism-Aware Models.” In: *Genetics* 212.4, pp. 1321–1336. DOI: `10.1534/genetics.119.302074`.
- Borges, Rui and Carolin Kosiol (2020). “Consistency and identifiability of the polymorphism-aware phylogenetic models.” In: *Journal of Theoretical Biology* 486, p. 110074. DOI: `10.1016/j.jtbi.2019.110074`.
- De Maio, Nicola, Dominik Schrempf, and Carolin Kosiol (2015). “PoMo: An Allele Frequency-Based Approach for Species Tree Estimation.” In: *Systematic Biology* 64.6, pp. 1018–1031. DOI: `10.1093/sysbio/syv048`.
- Hervas, Sergi, Esteve Sanz, Sònia Casillas, John E Pool, and Antonio Barbadilla (2017). “Popfly: the Drosophila Population Genomics Browser.” In: *Bioinformatics* 33.17, pp. 2779–2780. DOI: `10.1093/bioinformatics/btx301`.

Bibliography II

- Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear (2020). "IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era." In: *Molecular Biology and Evolution* 37.5. Ed. by Emma Teeling, pp. 1530–1534. DOI: `10.1093/molbev/msaa015`.
- Schrempf, Dominik, Bui Quang Minh, Nicola De Maio, Arndt von Haeseler, and Carolin Kosiol (2016). "Reversible polymorphism-aware phylogenetic models and their application to tree inference." In: *Journal of Theoretical Biology* 407, pp. 362–370. DOI: `10.1016/j.jtbi.2016.07.042`.
- Schrempf, Dominik and Asger Hobolth (2017). "An alternative derivation of the stationary distribution of the multivariate neutral Wright–Fisher model for low mutation rates with a view to mutation rate estimation from site frequency data." In: *Theoretical Population Biology* 114, pp. 88–94. DOI: `10.1016/j.tpb.2016.12.001`.

Bibliography III

Schrempf, Dominik, Bui Quang Minh, Arndt von Haeseler, and Carolin Kosiol (2019). “Polymorphism-Aware Species Trees with Advanced Mutation Models, Bootstrap, and Rate Heterogeneity.” In: *Molecular Biology and Evolution* 36.6. Ed. by Naruya Saitou, pp. 1294–1301. DOI: `10.1093/molbev/msz043`.