Workshop

Last update: Aug 8, 2019, Contributors: Minh Bui

IQ-TREE Workshop Tutorial (Woods Hole 2019)

Table of Contents *generated with DocToc (https://github.com/thlorenz/doctoc)*

If you haven't installed IQ-TREE, please download (http://www.iqtree.org/#download) and install (../doc/Quickstart) the binary for your platform. For the next steps, the folder containing your `iqtree` executable should be added to your PATH enviroment variable so that IQ-TREE can be invoked by simply entering `iqtree` at the command-line. Alternatively, you can also copy `iqtree` binary into your system search.

Run the command

```
iqtree
```

should display something like this to the screen:

```
IQ-TREE multicore version 1.6.11 for Mac OS X 64-bit built Jun  6 2019
Developed by Bui Quang Minh, Nguyen Lam Tung, Olga Chernomor,
Heiko Schmidt, Dominik Schrempf, Michael Woodhams.
```

1) Input data

We will use a Turtle data set to demonstrate the use of IQ-TREE throughout the workshop. Please download the following input files:

- turtle.fa (data/turtle.fa): The DNA alignment (in FASTA format), which is a subset of the original Turtle data set used to assess the phylogenetic position of Turtle relative to Crocodile and Bird (Chiari et al., 2012 (https://doi.org/10.1186/1741-7007-10-65)).

- turtle.nex (data/turtle.nex): The partition file (in NEXUS format) defining 29 genes, which are a subset of the published 248 genes (Chiari et al., 2012 (https://doi.org/10.1186/1741-7007-10-65)).

> **QUESTIONS:**
>
> - View the alignment in Jalview or your favourite alignment viewer.
> - Can you identify the gene boundary from the viewer? Does it roughly match the partition file?
> - Is there missing data? Which taxa seem to have most missing data?
> - Do you think if missing data can be problematic?

2) Inferring the first phylogeny

You can now start to reconstruct a maximum-likelihood (ML) tree for the Turtle data set (assuming that you are in the same folder where the alignment is stored):

```
iqtree -s turtle.fa -bb 1000 -nt AUTO
```

Options explained:

- `-s turtle.fa` to specify the input alignment as `turtle.fa` .
- `-bb 1000` to specify 1000 replicates for the ultrafast bootstrap (Minh et al., 2013 (https://doi.org/10.1093/molbev/mst024)).
- `-nt AUTO` to determine the best number of CPU cores to speed up the analysis.

This simple command will perform three important steps in one go:

1. Select best-fit model using ModelFinder (Kalyaanamoorthy et al., 2017 (https://doi.org/10.1038/nmeth.4285)).
2. Reconstruct the ML tree using the IQ-TREE search algorithm (Nguyen et al., 2015 (https://doi.org/10.1093/molbev/msu300)).
3. Assess branch supports using the ultrafast bootstrap - UFBoot (Minh et al., 2013 (https://doi.org/10.1093/molbev/mst024)).

Once the run is done, IQ-TREE will write several output files including:

- `turtle.fa.iqtree` : the main report file that is self-readable. You should look at this file to see the computational results. It also contains a textual representation of the final tree.
- `turtle.fa.treefile` : the ML tree in NEWICK format, which can be visualized in FigTree or any other tree viewer program.
- `turtle.fa.log` : log file of the entire run (also printed on the screen).
- `turtle.fa.ckp.gz` : checkpoint file used to resume an interrupted analysis.
- And a few other files.

QUESTIONS:

- Look at the report file `turtle.fa.iqtree` .
- What is the best-fit model? What do you know about this model?
- Visualise the tree `turtle.fa.treefile` in FigTree.
- Compare the tree with the published tree (Chiari et al., 2012 (https://doi.org/10.1186/1741-7007-10-65)). Are they the same or different?
- If different, where are the difference(s)?
- Look at the boostrap supports. Which branch(es) have a low support?

3) Applying partition model

We now perform a partition model analysis (Chernomor et al., 2016 (https://doi.org/10.1093/sysbio/syw037)), where one allows each partition to have its own model:

```
iqtree -s turtle.fa -spp turtle.nex -bb 1000 -nt AUTO
```

Options explained:

- `-spp turtle.nex` to specify an *edge-linked proportional* partition model (Chernomor et al., 2016 (https://doi.org/10.1093/sysbio/syw037)). That means, there is one set of branch lengths. But each partition can have proportionally shorter or longer tree length, representing slow or fast evolutionary rate, respectively.

QUESTIONS:

- Look at the report file `turtle.nex.iqtree` . What are the lowest- and highest-evolving genes?
- Compare the AIC/AICc/BIC score of partition model versus un-partition model done above. Which model is better?
- Visualise the tree `turtle.nex.treefile` in Figtree and compare it with the tree from the un-partitioned model. Are they the same or different? If different, where is the difference? Which tree agrees with the published tree (Chiari et al., 2012 (https://doi.org/10.1186/1741-7007-10-65))?
- Look at the boostrap supports. Which branch(es) have a low support?

4) Choosing the best partitioning scheme

We now perform the PartitionFinder algorithm (Lanfear et al., 2012 (https://doi.org/10.1093/molbev/mss020)) that tries to merge partitions to reduce the potential over-parameterization:

```
iqtree -s turtle.fa -spp turtle.nex -bb 1000 -nt AUTO -m MFP+MERGE -rcluster 10 -pre tu
rtle.merge
```

Options explained:

- `-m MFP+MERGE` to perform PartitionFinder followed by tree reconstruction.
- `-rcluster 10` to reduce computations by only examining the top 10% partitioning schemes using the

relaxed clustering algorithm (Lanfear et al., 2014 (https://doi.org/10.1186/1471-2148-14-82)).

- `-pre turtle.merge` to set the prefix for all output files as `turtle.merge.*`. This is to avoid overwriting outputs from the previous analysis.

QUESTIONS:

- Look at the report file `turtle.merge.iqtree`. How many partitions do we have now?
- Look at the AIC/AICc/BIC scores. Is it better or worse than those of the un-partition and partition models done previously?
- How does the tree look like now? How high/low are the bootstrap supports?

5) Tree topology tests

We now want to know whether the trees inferred for the Turtle data set have significantly different log-likelihoods or not. This can be conducted with the SH test (Shimodaira and Hasegawa, 1999 (https://doi.org/10.1093/oxfordjournals.molbev.a026201)), or expected likelihood weights (Strimmer and Rambaut, 2002 (https://doi.org/10.1098/rspb.2001.1862)).

First, concatenate the trees constructed by single and partition models into one file:

For Linux/MacOS:

```
cat turtle.fa.treefile turtle.nex.treefile >turtle.trees
```

For Windows:

```
type turtle.fa.treefile turtle.nex.treefile >turtle.trees
```

Now pass this file into IQ-TREE via `-z` option:

```
iqtree -s turtle.fa -spp turtle.nex.best_scheme.nex -z turtle.trees -zb 1000 -n 0 -wpl
-pre turtle.test
```

Options explained:

- `-spp turtle.nex.best_scheme.nex` to provide the partition model found previously to avoid running ModelFinder again.
- `-z turtle.trees` to input a set of trees.
- `-zb 1000` to specify 1000 replicates for *approximate* boostrap for tree topology tests.
- `-n 0` to avoid tree search and just perform tree topology tests.
- `-wpl` to print partition-wise log likelihoods for both trees. This will be used later in the next section.
- `-pre turtle.test` to set the prefix for all output files as `turtle.test.*`.

QUESTIONS:

- Look at the report file `turtle.test.iqtree`. There is a new section called `USER TREES`.
- Do the two trees have significantly different log-likelihoods?

HINTS:

- The KH and SH tests return p-values, thus a tree is rejected if its p-value < 0.05 (marked with a `-` sign).
- bp-RELL and c-ELW return posterior weights which are **not** p-value. The weights sum up to 1 across the trees tested.

6) Concordance factors

So far we have assumed that gene trees and species tree are equal. However, it is well known that gene trees might be discordant. Therefore, we now want to quantify the agreement between gene trees and species tree in a so-called *concordance factor* (Minh et al., 2018 (https://doi.org/10.1101/487801)). This feature is only available in the beta version 1.7-beta, please download it from https://github.com/Cibiv/IQ-TREE/releases/ (https://github.com/Cibiv/IQ-TREE/releases/). We assume that the command `iqtree-beta` in this section links to the beta version.

You first need to compute the gene trees, one for each partition separately:

```
iqtree-beta -s turtle.fa -S turtle.nex -pre turtle.loci -nt 2
```

Options explained:

- `-S turtle.nex` to tell IQ-TREE to infer separate trees for every partition in `turtle.nex` . All output files are similar to a partition analysis, except that the tree `turtle.loci.treefile` now contains a set of gene trees.

Definitions:

- **Gene concordance factor (gCF)** is the percentage of *decisive* gene trees concordant with a particular branch of the species tree (0% <= gCF(b) <= 100%). gCF=0% means that branch *b* does not occur in any gene trees, whereas gCF=100% means that branch *b* occurs in every gene tree.
- **Site concordance factor (sCF)** is the percentage of *decisive* (parsimony informative) alignment sites supporting a particular branch of the species tree (~33% <= sCF(b) <= 100%). sCF<33% means that another discordant branch *b'* is more supported, whereas sCF=100% means that branch *b* is supported by all sites.
- **CAUTION** when gCF ~ 0% or sCF < 33%, even if boostrap supports are ~100%!
- **GREAT** when gCF and sCF > 50% (i.e., branch is supported by a majority of genes and sites).

You can now compute gCF and sCF for the tree inferred under the partition model:

```
iqtree-beta -t turtle.nex.treefile --gcf turtle.loci.treefile -s turtle.fa --scf 100
```

Options explained:

- `-t turtle.nex.treefile` to specify a species tree.
- `--gcf turtle.loci.treefile` to specify a gene-trees file.
- `--scf 100` to draw 100 random quartets when computing sCF.

Once finished this run will write several files:

- `turtle.nex.treefile.cf.tree` : tree file where branches are annotated with bootstrap/gCF/sCF values.
- `turtle.nex.treefile.cf.stat` : a table file with various statistics for every branch of the tree.

Similarly, you can compute gCF and sCF for the tree under unpartitioned model:

```
iqtree-beta -t turtle.fa.treefile --gcf turtle.loci.treefile -s turtle.fa --scf 100
```

QUESTIONS:

- Visualise `turtle.nex.treefile.cf.tree` in FigTree.
- How do gCF and sCF values look compared with bootstrap supports?
- Visualise `turtle.fa.treefile.cf.tree` . How do these values look like now on the contradicting branch?

7) Resampling partitions and sites

Instead of bootstrap resampling sites, it is recommended to resample partitions and then sites within resampled partitions (Hoang et al., 2018 (https://doi.org/10.1093/molbev/msx281)). This may help to reduce over-confident branch supports.

```
iqtree -s turtle.fa -spp turtle.nex -bb 1000 -nt AUTO -bsam GENESITE -pre turtle.bsam
```

Options explained:

- `-bsam GENESITE` to turn on resampling partition and sites strategy.
- `-pre turtle.bsam` to set the prefix for all output files as `turtle.bsam.*` . This is to avoid overwriting outputs from the previous analysis.

QUESTIONS:

- Is there any change in the tree topology?
- Do the bootstrap support values get smaller or larger?

8) Identifying most influential genes

Now we want to investigate the cause for such topological difference between trees inferred by single and partition model. One way is to identify genes contributing most phylogenetic signal towards one tree but not the other.

How can one do this? Well, we can look at the gene-wise log-likelihood (logL) differences between the two given trees T1 and T2. Those genes having the largest logL(T1)-logL(T2) will be in favor of T1. Whereas genes showing the largest logL(T2)-logL(T1) are favoring T2.

With the `-wpl` option done above, IQ-TREE will write partition-wise log-likelihoods into `turtle.test.partlh` file.

QUESTIONS:

- Import this file into MS Excel. Compute the partition wise log-likelihood differences between two trees.
- What are the two genes that most favor the tree inferred by single model?
- Have a look at the paper by (Brown and Thomson, 2016 (https://doi.org/10.1093/sysbio/syw101)). Compare the two genes you found with those from this paper. What is special about these two genes?

9) Wrapping up

FINAL QUESTION:

- Given all analyses done in this tutorial, which tree do you think is the true tree?

Software

IQ-TREE release notes (/release/)

GitHub repository (https://github.com/Cibiv/IQ-TREE)

Development team (/about/)

Download statistics (http://www.somsubhra.com/github-release-stats/?username=Cibiv&repository=IQ-TREE)

User support

Frequently asked questions (/doc/Frequently-Asked-Questions)

User documentation (/doc/)

IQ-TREE Google group (https://groups.google.com/d/forum/iqtree)

IQ-TREE online web service (http://iqtree.cibiv.univie.ac.at/)

Links

Research School of Computer Science, ANU (https://cs.anu.edu.au)

Center for Integrative Bioinformatics Vienna (http://www.cibiv.at)

Link Checker (http://validator.w3.org/check/referer)