HHSD

Daniel Kornai

2023

Abstract

Introduction

Short example

Explanation of parameters

seqfile

`seqfile` specifies the location of the PHYLIP formatted multiple sequence alignment (MSA). The path to the file can be relative to the directory of the control file, or absolute, but HHSD will always attempt to resolve the absolute path, before beginning the analysis.

The naming of the sequences within the MSA is crucial, and should be in the form

`sequence_id^individual_id`

(e.g. `seq_1^ant320`), or `^individual_id` (e.g. `^ant320`). See the demonstration datasets for examples.

Imapfile

`Imapfile` specifies the location of the Imap file used to map individuals to their respective populations/species in the starting delimitation.

Best practices

Guide tree

The guide tree used for HHSD analyses has a fundamental impact on the results. If a widely accepted guide tree is not available for the groups being investigated,

Mutation rates

Given the inconsistent behaviour of the gdi alone at low effective population sizes, analyses with HHSD greatly benefit from the additional information provided by mutation rates. Users should aim to use empirically derived mutation rates specific to the taxonomic groups being investigated. If such data is not available, multiple analyses should be conducted using a range of mutation rates estimated in closely related groups.

Hidden prior estimation of τ and θ

As in all Bayesian analyses, The BPP A00 procedure used for intra-model inference of numeric parameters requires that a prior distribution is specified. While migration rate priors must be supplied by the user, τ and θ priors, when not specified in the master control file will be automatically inferred from the sequence data. The procedure is identical to the established program minimalist bpp. If the user is concerned about this practice, they should conduct analyses under multiple different priors, and check for convergence.

Ensuring run-to-run reproducibility

Due to the increase in the number of parameters, reliable inference of migration rate (M) parameters in the MSC+M model requires substantially larger amounts of data, relative to a basic MSC model with only τ and θ parameters. Accordingly, the number of specified migration events should be kept to a minimum, and practitioners should aim to provide as many sequences as they can for populations involved in migration events. In situations where the phylogenetic information is insufficient, the inferred migration rates in the analyses will simply be the prior mean values. To validate that the current dataset is able to constrain the MSC+M model, users should conduct multiple replicate analyses with different migration rate priors.

Advanced features

Parameter override from the command line

Consider the case when multiple replicate runs are to be conducted under different seeds, or with different priors to ensure convergence. In such cases, all other parameters (except the output directory), will stay consistent among runs. To ensure that such analyses do not require the creation of unnecessary control files, the program features the capability to override master control file parameters from the command line.

The default command for running the Giraffe analysis would be:

```
python3 HHSD.py --mcfile /Test_MCFs/Giraffe_merge.txt
```

To override certain parameters, the `--mcfpor` (master control file parameter override) option is used, which is followed by parameter name - parameter value pairs, separated by commas:

```
python3 HHSD.py --mcfile /Test_MCFs/Giraffe_merge.txt --mcfpor seed = 123, migprior = 0.01 10
```

When running the with parameter overrides, the program will inform the user:

```
< Checking control file arguments... >

The following master control file parameters have been overridden via --mcfpor:
seed = 123
migprior = 0.01 10
```

This feature enables replicate analyses to be specified using a bash script. See the included `giraffe.sh` demonstration file for an example.