Taxonomic classification using `pplacer`, *clst*, and *clstutils*

Noah Hoffman

June 14, 2024

Contents

1 Introduction

This vignette assumes that you have already created a reference package, and have used it to run `pplacer` against an alignment of reference sequences. For instruction on performing the above operations, see the documentation for `pplacer` here: `http://matsen.fhcrc.org/pplacer/`

2 Input files

The following input is required (file names defined as variables in parentheses):

1. a reference package (`refpkg`)
2. a pplacer file created using the same reference package (`placefile`)
3. a file providing distances between nodes in the reference tree (`distfile`)

Note that `distfile` is generated from `placefile` using `placeutil` (distributed with `pplacer`).

```
> library(clstutils)
> expand <- function(fname){
+   orig.dir <- getwd()
+   destdir <- tempdir()
+   setwd(destdir)
+   archive <- system.file('extdata','vaginal_16s.refpkg.tar.gz', package='clstutils')
+   system(sprintf('tar --no-same-owner -xzf "%s"', archive))
+   setwd(orig.dir)
+   file.path(destdir, fname)
+ }
> refpkg <- expand('vaginal_16s.refpkg')
> placefile <- system.file('extdata','merged.json', package='clstutils')
> distfile <- system.file('extdata','merged.distmat.bz2', package='clstutils')
```

3 Reading the input

Classification requires a matrix representation of distances between "objects" being classified, in this case sequences in a phylogenetic tree. `treeDists` returns a list containing matrix representations of distances between internal and terminal edges (`$dists` and `$paths`), and `$dmat`, a square matrix of distances between terminal edges.

```
> treedists <- treeDists(distfile=distfile, placefile=placefile)
```

We also need a description of the taxonomy of the reference sequences. This is read from the reference package using `taxonomyFromRefpkg`. The `seqnames` argument ensures that the output is arranged in an order compatible with `treedists`. We indicate that the most specific rank that we want to consider is "species" using `lowest_rank`.

```
> taxdata <- taxonomyFromRefpkg(refpkg, seqnames=rownames(treedists$dmat), lowest_rank='species')
```

4 Classification of a single sequence

Given the distances and taxonomic information describing the reference tree, the only additional data required to perform classification is the position of a sequence placed onto a tree. At a minimum, this consists of a data.frame with columns `at`, `edge`, and `branch`. This data will be used to generate a vector of branch lengths between the query sequence and each of the reference sequences on the tree.

```
> placetab <- data.frame(at=49, edge=5.14909e-07, branch=5.14909e-07)
```

The function `classifyPlacements` is a wrapper around *clst*::`classifyIter`. The output is a `data.frame` describing the taxonomic assignment, along with a description of the confidence of the classification. See the man page for *clst*::`classify` for details on the output.

```
> cdata <- classifyPlacements(taxdata, treedists, placetab)
> cdata

  tax_id              tax_name    rank below above score match min median max
1   2702 Gardnerella vaginalis species    10     0  0.95     1   0   0.06 0.1
     d at
1 0.16 49
```