The microcontax package vignette

Lars Snipen

1 Introduction

There is no official taxonomy for prokaryotes. This means that when we are building and testing methods for taxonomic classification of bacteria and archaea, there is no real gold standard data set that we can use for validating the results. The dominating marker used to study the evolution of prokaryotes is the 16S rRNA gene. There are several public data repositories for such data, but their taxonomic assignments differ. In an attempt to overcome this problem, we collected millions of full-length classified 16S sequences from various sources, and extracted the subset with some degree of consensus on the classification. This is the ConTax (Consensus Taxonomy) data set, see Vinje *et al.* (2016) for details.

In this package you find a trimmed version of the full data set, named `contax.trim`. This is designed to be used as training data for classifiers trying to recognize prokaryotes based on the 16S sequence. You will also find a set of medoid sequences, i.e. one representative from each genus.

2 Full data set

The full data set is too large to be part of a CRAN package. It is instead found in an additional data package named `micrcontax.data`. This is available at `https://khliland.github.io/drat/`. Here is how you install this package, provided you have already installed `microcontax`:

```
> if (!requireNamespace("microcontax.data", quietly = TRUE)) {
+  install.packages("microcontax.data")
+ }
```

You may now load the data set named `contax.full` as

```
> data("contax.full", package="microcontax.data")
```