The microseq package vignette

Lars Snipen and Kristian Hovde Liland

1 Using dplyr and stringr

An idea behind this package is to keep sequence data in the generic data structures in R instead of creating new and complex data types. This makes it possible to use the power of standard data manipulation tools that R-users are familiar with.

Both FASTA and FASTQ files are read into tables, and sequences are stored as texts. This makes it straightforward to use all the tools available in packages like dplyr and stringr, for data wrangling and string manipulations. Both input and output FASTA or FASTQ files may be gzipped, no need for uncompressions.

Functions for findings ORFs or genes (findOrfs, findrRNA, findGenes) return results as GFF-formatted tables, i.e. a standard tibble with either texts or numbers in the columns.

Many bioinformatic softwares produces results as tables, if you let them. Reading, wrangling and plotting data in tables is what R does best!

2 External software

Some functions in this package calls upons external software that must be available on the system. Some of these are 'installed' by simply downloading a binary executable that you put somewhere proper on your computer. To make such programs visible to R, you typically need to update your PATH environment variable, to specify where these executables are located. Try it out, and use google for help!

2.1 Software muscle

The functions *msalign()* and *muscle()* uses the free software muscle (https://www.drive5.com/muscle/). From the website you download (and unzip) an executable. NB! Change its name to muscle, no more and no less (i.e. no version numbers etc). In the R console the command

```
> system("muscle -h")
```

should produce some sensible output.

2.2 Software barrnap

The functions *findrRNA()* uses the free software `barrnap` (https://github.com/tseemann/barrnap). The GitHub site explains how to install. In the R console the command

```
> system("barrnap -h")
```

should produce some sensible output.

2.3 Software prodigal

The functions *findGenes()* uses the free software `prodigal` (https://github.com/hyattpd/Prodigal). The GitHub site explains how to install. In the R console the command

```
> system("prodigal -h")
```

should produce some sensible output.