The TargetSearchData Package

Álvaro Cuadros-Inostroza

October 24, 2025

Abstract

This package contains exemplary files for the package *TargetSearch*. It includes raw NetCDF files from an *E. coli* salt stress experiment, extracted peak list for each CDF file, a sample description file, a metabolite reference library, and a retention index marker definition file.

1 Package Contents

The *TargetSearchData* package contains as subset of metabolite data collected from a *E. coli* salt stress experiment. This was part of a stress response study described in Jozefczuk et al. [2010]. The samples were measured in an Agilent 6890 series GC system with a 7683 series autosampler injector coupled to a Leco Pegasus 2 time-of-flight mass spectrometer. For further details, please refer to said publication.

In addition, this package **only provides** a couple of simple *helper* functions; these are intended to be used by *TargetSearch* in its examples. This is discussed in the following section.

The following subsections describe each type of file in detail.

1.1 Sample File

Tab-delimite text file of sample metadata. This provides a list of raw `CDF` files, their respective measurement day (MD), and the time point of the salt stress experiment. Samples with the same time point correspond to biological replicates. Note that the time point is given in arbitrary units, more precisely, 1 is the first sampling time, 3 is the third one, and so on.

Note that the measurement day corresponds with the first four digits of the chromatogram file name.

samples.txt

CDF_FILE	MEASUREMENT_DAY	TIME_POINT
7235eg08.cdf	7235	1
7235eg11.cdf	7235	1
7235eg26.cdf	7235	1
7235eg04.cdf	7235	3
7235eg30.cdf	7235	3
...		

1.2 CDF files

The `CDF` files, also known as `NetCDF`, are a trimmed down version of the original, *baseline corrected*, files that were experted by *LECO Pegasus* in the context of the *E. coli* salt stress experiment. The baseline correction was performed by *LECO* using default parameters. For information on the experiment, please refer to Jozefczuk et al. [2010].

In particular, the retention time was bounded between 200 and 400 seconds, while the mass-over-charge ratio (m/z) between 85 and 320 daltons. This was done in order to reduce the package size.

The file names follow a systematic nomenclature: `ydddaann.cdf`, where `y` is the last digit of the year (starting from 2000), `ddd` is the day of the year (from 1 to 365), `aa` is an arbitrary code (originally this was connected to a specific mass spectrometer), and `nn` is the measurement order of the sample in the specified year and day. Together, the part `yddd` correspond to the measurement day.

1.3 RI files

For each `CDF` file there is a corresponding retention time corrected peak list file, the so-called `RI` files. These are tab-delimited text files containing the retention time, retention index, and spectra. Each spectrum is a list of *m/z* and intensities separated by colons (:).

The text file format is the original format used in early *TargetSearch* version. There is also a binary format which is used by default and is designed for fast reading. These files are not part of this package, however (see *TargetSearch* documentation).

The file name convention is to prefix each `CDF` file with `RI_` and change the file extension from `cdf` to `txt` (or `dat` for the binary format).

1.4 Library File

This is tab-delimited text file with the list of metabolite targets (library) to search in the chromatograms. Each row is a metabolite, while columns are the metabolite name, retention index (RI), time deviation (`Win_1`), selective masses (`SEL_MASS`), and the metabolite spectrum.

The time deviation is specified in the first column (`Win_1`) and represents the first search window. See *TargetSearch* vignette for details.

library.txt

```
Name            RI      Win_1  SEL_MASS             SPECTRUM
Pyruvic acid    222767  4000   89;115;158;174;189   85:7 86:14 87:7 88:5 ...
Glycine (2TMS)  228554  4000   86;102;147;176;204   86:26 87:19 88:8 89:4 ...
Valine          271500  2000   100;144;156;218;246  85:8 86:14 87:6 88:5 ...
Glycerol (3TMS) 292183  2000   103;117;205;293      85:14 86:2 87:16 88:13 ...
Leucine         306800  1500   102;158;232;260      158:999 159:148 160:45 ...
...
```

1.5 Retention Time Correction File

This is tab-delimited text file of retention time markers. The markers are fatty acid methyl esthers (FAME) which elute evenly distributed on the retention time dimension and have a characteristic *m/z* value.

rimLimits.txt

```
LowerLimit UpperLimit RIstandard
       230        280     262320
       290        340     323120
       350        400     381020
```

A fourth column called *Mass* can be specified if the *m/z* value is **not** the default of 87. See `ImportFameSettings` documentation from *TargetSearch*.

2 Helper functions

Simple functions are provided to return the absolute file path of the files contained in *TargetSearchData*. These function are used frequently in *TargetSearch* examples as a shorthand to function calls such as `find.package()` and `file.path()`. The functions also perform some basic checks internally. All these functions have a `tsd_` prefix. The following shows a brief description and examples of these functions.

The function `tsd_path()` is a function to locate *TargetSearchData* installation path, where the example files are found. It's built on `find.package()`.

```
> path <- tsd_path()
> path
```

```
[1] "/tmp/RtmpSU4XJ1/Rinst1a0ccc25a6a446/TargetSearchData"
```

The function `tsd_data_path()` returns the path where the example files are. They are in a subdirectory called `gc-ms-data`, which can be passed as argument, though this is not needed.

```
> path <- tsd_data_path()
> dir(path)
```

```
 [1] "7235eg04.cdf"    "7235eg06.cdf"    "7235eg07.cdf"    "7235eg08.cdf"
 [5] "7235eg09.cdf"    "7235eg11.cdf"    "7235eg12.cdf"    "7235eg15.cdf"
 [9] "7235eg20.cdf"    "7235eg21.cdf"    "7235eg22.cdf"    "7235eg25.cdf"
[13] "7235eg26.cdf"    "7235eg30.cdf"    "7235eg32.cdf"    "RI_7235eg04.txt"
[17] "RI_7235eg06.txt" "RI_7235eg07.txt" "RI_7235eg08.txt" "RI_7235eg09.txt"
[21] "RI_7235eg11.txt" "RI_7235eg12.txt" "RI_7235eg15.txt" "RI_7235eg20.txt"
[25] "RI_7235eg21.txt" "RI_7235eg22.txt" "RI_7235eg25.txt" "RI_7235eg26.txt"
[29] "RI_7235eg30.txt" "RI_7235eg32.txt" "library.txt"     "rimLimits.txt"
[33] "samples.txt"
```

The function `tsd_file_path()` returns of path of one or more files contained in the directory `gc-ms-data`. It takes a character vector as argument. If the file does not exist, then it raises an error.

```
> tsd_file_path(c('samples.txt','library.txt'))
```

```
[1] "/tmp/RtmpSU4XJ1/Rinst1a0ccc25a6a446/TargetSearchData/gc-ms-data/samples.txt"
[2] "/tmp/RtmpSU4XJ1/Rinst1a0ccc25a6a446/TargetSearchData/gc-ms-data/library.txt"
```

Finally, the functions `tsd_cdffiles` and `tsd_rifiles` return the list of `CDF` and `RI` files contained in the data path respectively.

```
> cdf <- tsd_cdffiles()
> ri <- tsd_rifiles()
```

3 Session Info

Output of `sessionInfo()` on the system on which this document was compiled.

```
> sessionInfo()
```

```
R Under development (unstable) (2025-10-20 r88955)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.3 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.23-bioc/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0  LAPACK version 3.12.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB              LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] TargetSearchData_1.47.0

loaded via a namespace (and not attached):
[1] compiler_4.6.0 tools_4.6.0
```

References

Szymon Jozefczuk, Sebastian Klie, Gareth Catchpole, Jedrzej Szymanski, Alvaro Cuadros-Inostroza, Dirk Steinhauser, Joachim Selbig, and Lothar Willmitzer. Metabolomic and transcriptomic stress response of escherichia coli. *Molecular Systems Biology*, 6(1):364, 2010. doi: https://doi.org/10.1038/msb.2010.18.