GEOsubmission

Alexandre Kuhn (*kuhnam@mail.nih.gov*)

May 3, 2016

1 Summary

The goal of *GEOsubmission* is to ease the submission of microarray datasets to the GEO repository. It generates a single file (SOFT format) containing sample information and gene expression values. This file can then be uploaded to GEO in a single step.

2 Introduction

The rate of microarray data deposition in public repository is low [Ochsner et al., 2008]. *GEOsubmission* provides a simple and quick way to create a dataset submission for deposition at GEO (`http://www.ncbi.nlm.nih.gov/geo/`).

3 Principle

Submitting a microarray dataset to a public repository generally implies to gather information on the samples (including how they were processed) and upload this information along with the microarray data to a repository. Having both sample information and gene expression data in a single file greatly eases the submission process. GEO accepts SOFT files (`http://www.ncbi.nlm.nih.gov/geo/info/soft2.html`), a file format that can be used to describe samples and expression values in a single text file.

GEOsubmission contains a function (`microarray2soft`) that generates a SOFT file. Sample information is gathered from user-provided text files, which are then bundled together with the corresponding expression values in the SOFT file. Microarray expression data is provided as a single tab-delimited text file (with rows corresponding to probes and columns to samples). In the case of Affymetrix microarrays, RMA-normalized expression can be calculated directly from the CEL files, i.e. without providing a separate text file for the expression values.

Sample information is provided in two text files. The first one describes each sample in the dataset. The second one provides information on the dataset itself (named a "series" by GEO). `microarray2soft` performs consistency checks on sample and series information as well as makes sure that they match the corresponding micorraray raw data files. Once the SOFT file is created, it can be compressed together with the raw data microarray data in a zip file. This zip file can be used for a single-step deposition (named "direct deposit") at GEO.

4 Example usage

Consider an example experiment where we assayed gene expression in neuronal cultures using Affymetrix microarrays. Let us assume that this dataset (named "neuronalCultures") contains two samples (named "1" and "2"), corresponding to CEL files "sample1.CEL" and "sample2.CEL", respectively. Let us assume further that sample and series information is contained in the files named "sampleInfo.txt" and "seriesInfo.txt" respectively (in the current directory). You can use the following code to create a SOFT file named "example.soft" in the current directory:

```
> library(GEOsubmission)

> sampleID<-c('1','2')
> seriesName<-'neuronalCultures'

> microarray2soft(sampleID,'sampleInfo.txt',seriesName,'seriesInfo.txt',
+ softname='mydata.soft')
>
```

Note that this first example cannot be run because we only provide dummy, not valid CEL files with this package.

The format and content of the sample and series files is given in the two corresponding example files provided with this package. Their content can be shown in R with the following commands (or by opening them; they are located in the "extdata" directory contained in the installation directory of *GEOsubmission*):

```
> dataDirectory<-system.file('extdata',package='GEOsubmission')

> read.delim(file.path(dataDirectory,'sampleInfo.txt'))
> read.delim(file.path(dataDirectory,'seriesInfo.txt'))
```

Alternatively, for instance in the case of a microarray experiment using a different platform than Affymetrix, we can provide the gene expression values (for inclusion in the SOFT file) in a separate text file. If this is the case, `microarray2soft` will only check that the raw microarray data files given in 'sampleInfo.txt' actually exists and will use the separate text file as the source of expression values. In our neuronal culture example, if we wished to use the expression values in the file "expressionNormalized.txt" (instead of calculating normalized expression from the microarray data files), we can use the following. We first specify a directory and a file to write the generated example SOFT file out to:

```
> soft_example_fullpath<-tempfile(pattern='soft_example')
> soft_example_name<-basename(soft_example_fullpath)
> soft_example_dir<-dirname(soft_example_fullpath)
```

and then generate and write the SOFT file with:

```
> microarray2soft(sampleID,'sampleInfo.txt',seriesName,'seriesInfo.txt',
+ datadir=dataDirectory,writedir=soft_example_dir,
+ softname=soft_example_name,expressionmatrix='expressionNormalized.txt')
>
```

The generated SOFT file looks like this:

```
> readLines(soft_example_fullpath)
```

```
 [1] "^SAMPLE = 1"
 [2] "!Sample_title = sample1"
 [3] "!Sample_supplementary_file = sample1.CEL"
 [4] "!Sample_source_name = Brain"
 [5] "!Sample_organism = Rattus norvegicus"
 [6] "!Sample_characteristics = Primary neuronal culture"
 [7] "!Sample_molecule = Total RNA"
 [8] "!Sample_extract_protocol = TRIzol (Invitrogen) followed by RNeasy column cleanup (Qi
 [9] "!Sample_label = Biotin"
[10] "!Sample_label_protocol = Affymetrix GeneChip\xae IVT Labeling Kit, according to manu
[11] "!Sample_hyb_protocol = Affymetrix Eukaryotic Target Hybridization protocol (GeneChip
[12] "!Sample_scan_protocol = Affymetrix\xae GeneChip\xae Scanner 3000 with GCOS software
[13] "!Sample_data_processing = Probe set summarization and normalization was performed by
[14] "!Sample_description = Primary culture from cerebral cortices of rat P1 pups"
[15] "!Sample_platform_id = GPL1355"
[16] "#ID_REF ="
[17] "#VALUE = RMA-calculated signal intensity"
[18] "!Sample_table_begin"
[19] "ID_REF\tVALUE"
[20] "probe1\t6"
[21] "probe2\t5"
[22] "!Sample_table_end"
[23] "^SAMPLE = 2"
[24] "!Sample_title = sample2"
[25] "!Sample_supplementary_file = sample2.CEL"
[26] "!Sample_source_name = Brain"
[27] "!Sample_organism = Rattus norvegicus"
[28] "!Sample_characteristics = Primary neuronal culture"
[29] "!Sample_molecule = Total RNA"
[30] "!Sample_extract_protocol = TRIzol (Invitrogen) followed by RNeasy column cleanup (Qi
[31] "!Sample_label = Biotin"
[32] "!Sample_label_protocol = Affymetrix GeneChip\xae IVT Labeling Kit, according to manu
[33] "!Sample_hyb_protocol = Affymetrix Eukaryotic Target Hybridization protocol (GeneChip
[34] "!Sample_scan_protocol = Affymetrix\xae GeneChip\xae Scanner 3000 with GCOS software
[35] "!Sample_data_processing = Probe set summarization and normalization was performed by
[36] "!Sample_description = Primary culture from cerebral cortices of rat P1 pups"
[37] "!Sample_platform_id = GPL1355"
[38] "#ID_REF ="
[39] "#VALUE = RMA-calculated signal intensity"
[40] "!Sample_table_begin"
[41] "ID_REF\tVALUE"
[42] "probe1\t7"
[43] "probe2\t4"
[44] "!Sample_table_end"
[45] "^SERIES = neuronalCultures"
[46] "!Series_title = Gene expression from primary neuronal cultures."
[47] "!Series_summary = Gene expression from primary neuronal cultures."
```

```
[48] "!Series_type = Primary cell cultures"
[49] "!Series_overall_design = Primary neuronal cultures (2 biological replicates)."
[50] "!Series_contributor = John,Smith"
[51] "!Series_contributor =  William,Ford"
[52] "!Series_sample_id = 1"
[53] "!Series_sample_id = 2"
```

The format of the file containing expression values is shown with the example file "expressionNormalized.txt" that is contained in this package. It can be output to the R console with the following command (it resides in the "extdata" directory of the package installation directory):

```
> read.delim(file.path(dataDirectory,'expressionNormalized.txt'))
```

The SOFT file can also be written to the standard output with:

```
> microarray2soft(c('1','2'),'sampleInfo.txt',seriesName,'seriesInfo.txt',
+ datadir=dataDirectory,softname='',expressionmatrix='expressionNormalized.txt',
+ verbose=FALSE)
>
```

More detailed information on the input arguments of `microarray2soft` are given in the help file that can be accessed by typing "?microarray2soft" at the R prompt.

5 Session Information

The version number of R and packages loaded for generating the vignette were:

```
R version 3.3.0 (2016-05-03)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.4 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        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

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

other attached packages:
[1] GEOsubmission_1.24.0

loaded via a namespace (and not attached):
[1] zlibbioc_1.18.0       BiocInstaller_1.22.0  parallel_3.3.0
[4] tools_3.3.0           affy_1.50.0           affyio_1.42.0
[7] Biobase_2.32.0        preprocessCore_1.34.0 BiocGenerics_0.18.0
```

References

Scott Ochsner, David Steffen, Christian J Stoeckert, and Neil J McKenna. Much room for improvement in deposition rates of expression microarray datasets. *Nat. Methods*, 5(12):991, 2008. doi: 10.1038/nmeth1208-991. URL `http://www.nature.com/nmeth/journal/v5/n12/suppinfo/nmeth1208-991_S1.html`.