Introduction to the Bioconductor marray package : Input component

Yee Hwa Yang[1] and Sandrine Dudoit[2]

August 25, 2004

1. Department of Medicine, University of California, San Francisco, jean@biostat.berkeley.edu
2. Division of Biostatistics, University of California, Berkeley,
http://www.stat.berkeley.edu/~sandrine

Contents

1 Overview

This document provides a tutorial for the **data input** component of the `marray` package. This is similar to the previous `marrayInput` package which has now been combined with the suite of other four packages for diagnostic plots and normalization of cDNA microarray data. This package relies on object–oriented class/method mechanism, provided by the R `methods` package, to allow efficient and systematic representation and manipulation of microarray data.
This vignette describeds functionality for reading microarray data into R, such as intensity data from image processing output files (e.g. `.spot` and `.gpr` files for the `Spot` and `GenePix` packages, respectively) and textual information on probes and targets (e.g. from gal files and god lists). A `tcltk` widget is supplied to facilitate and automate data input and the creation of microarray specific R objects for storing these data.

2 Getting started

Installing the package. To install the *marray* package for Windows operating systems, first start R and make sure you are connected to the internet. Next, select "**Packages**" from the menu and

click on " **Install package(s) from Bioconductor...**". Lastly, select *marray* from the pop–up windows and click click on "**OK**". You will find more detailed installation instructions on the Bioconductor web site.
Loading the package. To load the `marray` package in your R session, type `library(marray)`.
Help files. As with any R package, detailed information on functions, classes and methods can be obtained in the help files. For instance, to view the help file for the function `read.GenePix` in a browser, use `help.start()` followed by `? read.GenePix`.

A quick start guide on this package are given in the `marray` document of the `inst/doc` directory.

3 Case study: Swirl zebrafish microarray experiment

We demonstrate the functionality of this collection of R packages using gene expression data from the Swirl zebrafish experiment. These data were provided by Katrin Wuennenberg–Stapleton from the Ngai Lab at UC Berkeley. (The swirl embryos for this experiment were provided by David Kimelman and David Raible at the University of Washington.) This experiment was carried out using zebrafish as a model organism to study early development in vertebrates. Swirl is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis. Ventral fates such as blood are reduced, whereas dorsal structures such as somites and notochord are expanded. A goal of the Swirl experiment is to identify genes with altered expression in the swirl mutant compared to wild–type zebrafish. Two sets of dye–swap experiments were performed, for a total of four replicate hybridizations. For each of these hybridizations, target cDNA from the swirl mutant was labeled using one of the Cy3 or Cy5 dyes and the target cDNA wild–type mutant was labeled using the other dye. Target cDNA was hybridized to microarrays containing 8,448 cDNA probes, including 768 controls spots (e.g. negative, positive, and normalization controls spots). Microarrays were printed using 4×4 print–tips and are thus partitioned into a 4×4 grid matrix. Each grid consists of a 22×24 spot matrix that was printed with a single print–tip. Here, spot row and plate coordinates should coincide, as each row of spots corresponds to probe sequences from the same 384 well–plate.

Each of the four hybridizations produced a pair of 16–bit images, which were processed using the image analysis software package `Spot` (Buckley, 2000; Yang et al., 2002). Raw images of the Cy3 and Cy5 fluorescence intensities for all four hybridizations are available at `http://fgl.lsa.berkeley.edu/Swirl/index.html`. The dataset includes four output files `swirl.1.spot`, `swirl.2.spot`, `swirl.3.spot`, and `swirl.4.spot` from the `Spot` package. Each of these files contains 8,448 rows and 30 columns; rows correspond to spots and columns to different statistics from the `Spot` image analysis output. The file `fish.gal` is a gal file generated by the `GenePix` program; it contains information on individual probe sequences, such as gene names, spot ID, spot coordinates. Hybridization information for the mutant and wild–type target samples is stored in `SwirlSample.txt`. All fluorescence intensity data from processed images are included in the `marrayInput` package (see Section 4 for greater details).

To load the swirl dataset, use `data(swirl)`, and to view a description of the experiments and data, type `? swirl`. Below, we give step–by–step instructions for reading the swirl data into R. For

convenience, we have also stored the resuls in the object `swirl` of class `marrayRaw`.

```
> library(marray)
> data(swirl)
> swirl

An object of class "marrayRaw"
@maRf
     swirl.1.spot swirl.2.spot swirl.3.spot swirl.4.spot
[1,]    19538.470    16138.720    2895.1600   14054.5400
[2,]    23619.820    17247.670    2976.6230   20112.2600
[3,]    21579.950    17317.150    2735.6190   12945.8500
[4,]     8905.143     6794.381     318.9524     524.0476
[5,]     8676.095     6043.542     780.6667     304.6190
8443 more rows ...

@maGf
     swirl.1.spot swirl.2.spot swirl.3.spot swirl.4.spot
[1,]    22028.260    19278.770    2727.5600   19930.6500
[2,]    25613.200    21438.960    2787.0330   25426.5800
[3,]    22652.390    20386.470    2419.8810   16225.9500
[4,]     8929.286     6677.619     383.2381     786.9048
[5,]     8746.476     6576.292     901.0000     468.0476
8443 more rows ...

@maRb
     swirl.1.spot swirl.2.spot swirl.3.spot swirl.4.spot
[1,]          174          136           82           48
[2,]          174          133           82           48
[3,]          174          133           76           48
[4,]          163          105           61           48
[5,]          140          105           61           49
8443 more rows ...

@maGb
     [,1] [,2] [,3] [,4]
[1,]  182  175   86   97
[2,]  171  183   86   85
[3,]  153  183   86   85
[4,]  153  142   71   87
[5,]  153  142   71   87
8443 more rows ...

@maW
<0 x 0 matrix>
```

```
@maLayout
An object of class "marrayLayout"
@maNgr
[1] 4

@maNgc
[1] 4

@maNsr
[1] 22

@maNsc
[1] 24

@maNspots
[1] 8448

@maSub
[1] TRUE

@maPlate
factor(0)
Levels:

@maControls
[1] Control Control Control Control Control
Levels: Control N
8443 more elements ...

@maNotes
[1] "No Input File"


@maGnames
An object of class "marrayInfo"
@maLabels
[1] "geno1" "geno2" "geno3" "3XSSC" "3XSSC"
8443 more elements ...

@maInfo
     "ID" "Name"
1 control  geno1
2 control  geno2
3 control  geno3
4 control  3XSSC
```

```
5 control  3XSSC
8443 more rows ...

@maNotes
[1] "C:/GNU/R/rw1041/library/marrayInput/data/fish.gal"


@maTargets
An object of class "marrayInfo"
@maLabels
[1] "81" "82" "93" "94"

@maInfo
  # of slide        Names experiment Cy3 experiment Cy5      date comments
1         81 swirl.1.spot          swirl      wild type 2001/9/20       NA
2         82 swirl.2.spot      wild type          swirl 2001/9/20       NA
3         93 swirl.3.spot          swirl      wild type 2001/11/8       NA
4         94 swirl.4.spot      wild type          swirl 2001/11/8       NA

@maNotes
[1] "C:/GNU/R/rw1041/library/marrayInput/data/SwirlSample.txt"


@maNotes
[1] ""
```

4 Package marrayInput – Reading microarray data into R

We begin our analysis of microarray data with the fluorescence intensities produced by image processing of the microarray scanned images. These data are typically stored in tables whose rows correspond to the spotted probe sequences and columns to different spot statistics: e.g. grid row and column coordinates, spot row and column coordinates, red and green background and foreground intensities for different segmentation and background adjustment methods, spot morphology statistics, etc. For the `GenePix` image processing software, these are the `.gpr` files, and for `Spot`, these are the `.spot` files. We also consider probe and target textual information stored, for example, in `.gal` and `.gdl` (god list) files. The main functions in the `marrayInput` package are `read.marrayLayout`, `read.marrayInfo`, and `read.marrayRaw`, which create objects of classes `marrayLayout`, `marrayInfo`, and `marrayRaw`, respectively. Widgets are provided for each of these functions to facilitate data entry.

For the Swirl zebrafish experiment, textual information and fluorescence intensity data from processed images were included as part of the `marrayInput` package and can be accessed as follows, where `datadir` is the name of the R package sub–directory containing the data files.

```
> datadir <- system.file("swirldata", package = "marray")
> dir(datadir)
```

```
[1] "SwirlSample.txt" "fish.gal"        "swirl.1.spot"    "swirl.2.spot"
[5] "swirl.3.spot"    "swirl.4.spot"
```

4.1 Main input functions

Consider first the function `read.marrayLayout`, which may be used to read in and store information on the layout of spots in a batch of arrays. The main quantities are the dimensions of the grid and spot matrices. In addition, it is useful to keep track of information on the location and nature of control spots, and the print–tip–group and plate origin of the probes. The following command stores such layout information in the object `swirl.layout` of class `marrayLayout`. The location of the control spots is extracted from the fourth (`ctl.col=4`) column of the file `fish.gal`.

```
> swirl.layout <- read.marrayLayout(fname = file.path(datadir,
+     "fish.gal"), ngr = 4, ngc = 4, nsr = 22, nsc = 24, skip = 21,
+     ctl.col = 4)
> ctl <- rep("Control", maNspots(swirl.layout))
> ctl[maControls(swirl.layout) != "control"] <- "N"
> maControls(swirl.layout) <- factor(ctl)
> swirl.layout

An object of class "marrayLayout"
@maNgr
[1] 4

@maNgc
[1] 4

@maNsr
[1] 22

@maNsc
[1] 24

@maNspots
[1] 8448

@maSub
[1] TRUE

@maPlate
factor(0)
Levels:

@maControls
[1] Control Control Control Control Control
Levels: Control N
```

```
8443 more elements ...

@maNotes
[1] "/BIOCPkg/lib/marray/swirldata/fish.gal"
```

Objects of class `marrayInfo` may be used to store information on probe sequences and target samples. The following commands create such objects for the Swirl experiment, by reading in text files supplied by the experimenter.

```
> swirl.samples <- read.marrayInfo(file.path(datadir, "SwirlSample.txt"))
> swirl.samples

An object of class "marrayInfo"
@maLabels
[1] "81" "82" "93" "94"

@maInfo
  # of slide        Names experiment Cy3 experiment Cy5      date comments
1         81 swirl.1.spot          swirl      wild type 2001/9/20       NA
2         82 swirl.2.spot      wild type          swirl 2001/9/20       NA
3         93 swirl.3.spot          swirl      wild type 2001/11/8       NA
4         94 swirl.4.spot      wild type          swirl 2001/11/8       NA

@maNotes
[1] "/BIOCPkg/lib/marray/swirldata/SwirlSample.txt"

> swirl.gnames <- read.marrayInfo(file.path(datadir, "fish.gal"),
+     info.id = 4:5, labels = 5, skip = 21)
> swirl.gnames

An object of class "marrayInfo"
@maLabels
[1] "geno1" "geno2" "geno3" "3XSSC" "3XSSC"
8443 more elements ...

@maInfo
     "ID" "Name"
1 control  geno1
2 control  geno2
3 control  geno3
4 control  3XSSC
5 control  3XSSC
8443 more rows ...

@maNotes
[1] "/BIOCPkg/lib/marray/swirldata/fish.gal"
```

The function `read.marrayRaw` takes as its main argument a list of names for files containing the intensity data (e.g. `GenePix` output files `.gpr`). It also takes as arguments the names of already created layout, probe, and target description objects, e.g., `swirl.layout`, `swirl.gnames`, and `swirl.samples` for the Swirl experiment. The following commands read in all the `Spot` files residing in the `datadir` directory. The arguments further specify that the red and green foreground intensities are stored under the headings `Rmean` and `Gmean`, and that the red and green background intensities are store under the headings `morphR` and `morphG`, respectively.

```
> fnames <- dir(path = datadir, pattern = paste("*", "spot", sep = "."))
> swirl.raw <- read.marrayRaw(fnames, path = datadir, name.Gf = "Gmean",
+     name.Gb = "morphG", name.Rf = "Rmean", name.Rb = "morphR",
+     layout = swirl.layout, gnames = swirl.gnames, targets = swirl.samples)

[1] "in read.marrayraw"
[1] "/BIOCPkg/lib/marray/swirldata"
Reading ...  /BIOCPkg/lib/marray/swirldata/swirl.1.spot
Reading ...  /BIOCPkg/lib/marray/swirldata/swirl.2.spot
Reading ...  /BIOCPkg/lib/marray/swirldata/swirl.3.spot
Reading ...  /BIOCPkg/lib/marray/swirldata/swirl.4.spot
```

4.2 Widgets for input functions

To facilitate the creation of microarray data objects, each of these three input functions has a corresponding `tcltk` widget: `widget.marrayLayout`, `widget.marrayInfo`, and `widget.marrayRaw`. A screen–shot of the `marrayRaw` widget is shown in Figure 1; the command to launch the widget is as follows (here, `ext` specifies the image output file extension)

```
widget.marrayRaw(path=datadir, ext="spot")
```

4.3 Wrapper input functions

For users who prefer command line input for a specific class of image processing output files, we have defined three additional functions. The functions `read.Spot`, `read.GenePix`, and `read.SMD` automate the creation of `marrayRaw` objects from `Spot` and `GenePix` image analysis files, and from the Stanford Microarray Database (SMD) raw data files (`.xls`). The main arguments to these functions are a list of files and the directory path of the files. The following commands read two specific files from the `datadir` directory.

```
fnames <- dir(path=datadir,pattern=paste("*", "spot", sep="\."))[1:2]
swirl <- read.Spot(fnames, path=datadir,
          layout = swirl.layout,
          gnames = swirl.gnames,
          targets = swirl.samples)
```

Alternatively, without specifying any arguments, the functions `read.spot` and `read.GenePix` by default will read in all `Spot` or `GenePix` files within a current working directory. One has the option of setting the layout, probe, and target information manually at a later stage.

```
swirl <- read.Spot()
test.raw <- read.GenePix()
```

Note: Sweave. This document was generated using the `Sweave` function from the R *tools* package. The source file is in the `/inst/doc` directory of the package *marray*.

References

M. J. Buckley. *The Spot user's guide*. CSIRO Mathematical and Information Sciences, August 2000. `http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm`.

Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed. Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistic*, 11(1), 2002.

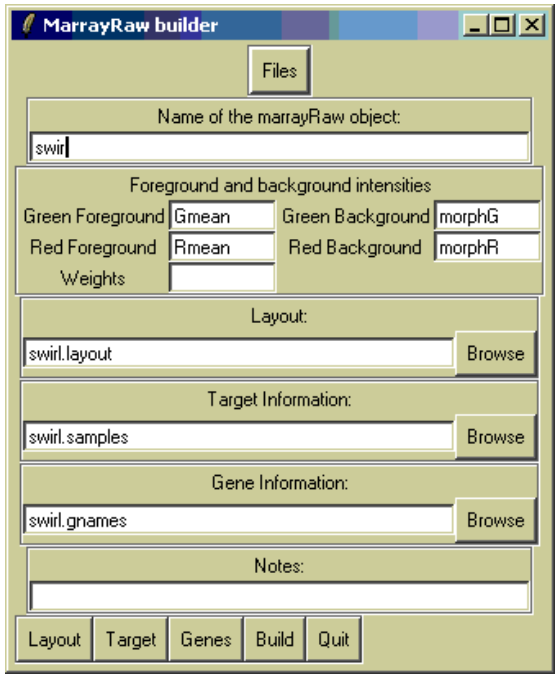

Figure 1: Screenshot of the widget for creating objects of class `marray` from image processing output files.