Introduction to the Bioconductor marrayInput package

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

May 13, 2004

1. Department of Statistics, University of California, Berkeley, yeehwa@stat.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 marrayInput package, which is part of a suite of four packages for diagnostic plots and normalization of cDNA microarray data. Like most Bioconductor packages, these four packages rely on the object–oriented class/method mechanism, provided by the R methods package, to allow efficient and systematic representation and manipulation of microarray data. marrayInput provides 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. The other three packages are

marrayClasses. This package contains basic class definitions and associated methods for pre– and post–normalization intensity data for batches of arrays.

marrayPlot. This package provides functions for diagnostic plots of microarray spot statistics, such as boxplots, scatter–plots, and spatial color images. Examination of diagnostic plots of intensity data is important in order to identify printing, hybridization, and scanning artifacts which can lead to biased inference concerning gene expression.

`marrayNorm`. This package implements robust adaptive location and scale normalization procedures, which correct for different types of dye biases (e.g. intensity, spatial, plate biases) and allow the use of control sequences spotted onto the array and possibly spiked into the mRNA samples. Normalization is needed to ensure that observed differences in intensities are indeed due to differential expression and not experimental artifacts; fluorescence intensities should therefore be normalized before any analysis which involves comparisons among genes within or between arrays.

2 Getting started

Installing the package. To install the `marrayInput` package for Windows operating systems, first download the file `marrayInput-snapshot.zip` from the Bioconductor website `http://www.bioconductor.org/packages/html/marrayInput.html`. Next, after starting R, from the menu select `Packages`, then `Install package from local zip file...`. Find and highlight the location of the zip file and click on `open`.

Loading the package. To load the `marrayInput` package in your R session, type `library(marrayInput)`.

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.marrayLayout` in a browser, use `help.start()` followed by `? read.marrayLayout`.

Microarray classes. The `marrayInput` packages relies on microarray class definitions in `marrayClasses`. You should also install this package and consult its vignette for more information.

Next. After reading your data into R, you can use the `marrayPlots` and `marrayNorm` packages for diagnostic plots and normalization, respectively.

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 `marrayInput`.

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(marrayInput)

Loading required package: marrayClasses
Welcome to Bioconductor
         Vignettes contain introductory material.  To view,
         simply type: openVignette()
         For details on reading vignettes, see
         the openVignette help page.

> data(swirl)
> swirl

Pre-normalization intensity data:         Object of class marrayRaw.

Number of arrays:         4 arrays.

A) Layout of spots on the array:
Array layout:         Object of class marrayLayout.

Total number of spots:                        8448
Dimensions of grid matrix:                4 rows by 4 cols
Dimensions of spot matrices:                22 rows by 24 cols

Currently working with a subset of 8448 spots.

Control spots:
There are   2 types of controls :
```

```
Control       N
    768    7680


Notes on layout:
No Input File

B) Samples hybridized to the array:
Object of class marrayInfo.

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

Number of labels:  4
Dimensions of maInfo matrix:  4  rows by  6  columns

Notes:
C:/GNU/R/rw1041/library/marrayInput/data/SwirlSample.txt

C) Summary statistics for log-ratio distribution:
              Min. 1st Qu. Median  Mean 3rd Qu. Max.
swirl.1.spot -2.73   -0.79  -0.58 -0.48   -0.29 4.42
swirl.2.spot -2.72   -0.15   0.03  0.03    0.21 2.35
swirl.3.spot -2.29   -0.75  -0.46 -0.42   -0.12 2.65
swirl.4.spot -3.21   -0.46  -0.26 -0.27   -0.06 2.90

D) Notes on intensity data:
```

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("data", package = "marrayInput")
> dir(datadir)

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

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

Array layout:         Object of class marrayLayout.

Total number of spots:                        8448
Dimensions of grid matrix:                4 rows by 4 cols
Dimensions of spot matrices:                22 rows by 24 cols

Currently working with a subset of 8448 spots.

Control spots:
There are   2 types of controls :

Control       N
    768    7680


Notes on layout:
/homes/madman/R-1.9.0/library/marrayInput/data/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

Object of class marrayInfo.

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

Number of labels:  4
Dimensions of maInfo matrix:  4  rows by  6  columns

Notes:
/homes/madman/R-1.9.0/library/marrayInput/data/SwirlSample.txt

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

Object of class marrayInfo.

   maLabels    "ID" "Name"
1     geno1 control  geno1
2     geno2 control  geno2
3     geno3 control  geno3
4     3XSSC control  3XSSC
5     3XSSC control  3XSSC
6      EST1 control   EST1
7     geno1 control  geno1
8     geno2 control  geno2
9     geno3 control  geno3
10    3XSSC control  3XSSC
...

Number of labels:  8448
Dimensions of maInfo matrix:  8448  rows by  2  columns
```

```
Notes:
/homes/madman/R-1.9.0/library/marrayInput/data/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] 0
[1] "Reading /homes/madman/R-1.9.0/library/marrayInput/data/swirl.1.spot"
[1] 0
[1] "Reading /homes/madman/R-1.9.0/library/marrayInput/data/swirl.2.spot"
[1] 0
[1] "Reading /homes/madman/R-1.9.0/library/marrayInput/data/swirl.3.spot"
[1] 0
[1] "Reading /homes/madman/R-1.9.0/library/marrayInput/data/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()
slot(swirl, "layout") <- swirl.layout
slot(swirl, "gnames") <- swirl.gnames
```

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.

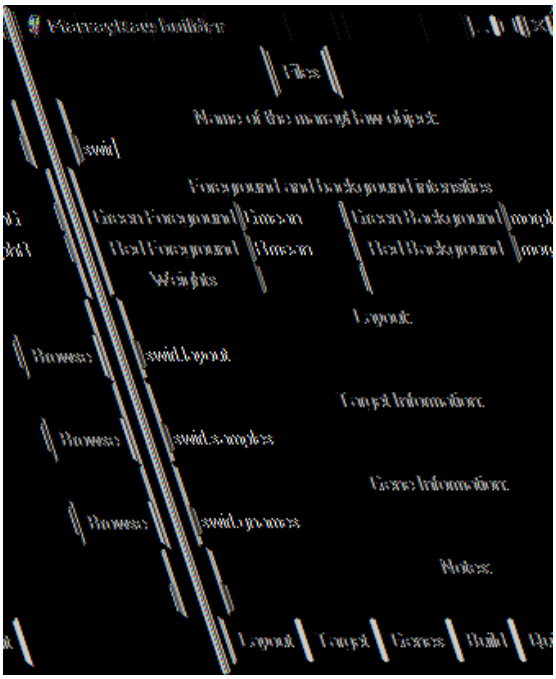

Figure 1: Screenshot of the widget for creating objects of class `marrayRaw` from image processing output files.