Multiple sample comparison in flow cytometry data with `flowMap`

Chiaowen Joyce Hsiao

Center for Bioinformatics and Computational Biology
University of Maryland, College Park

chsiao@umiacs.umd.edu

Modified: September 24th, 2013. Compiled: October 15, 2013

Contents

1 Introduction

Flow cytometry (FCM) is a powerful single-cell technology that provides high-throughput data on cellular features, such as size, complexity, and antibodies. The data is processed one sample at a time. Homegeneous groups of cells, also known as cell populations, are then identified via automatic or manual gating methods. These cell populations may vary in proprotion, shape or levels of a particular antibody markers between samples. A major step in the downstream analysis of flow cytometry data is to analyze cell population differences for phenotype comparisons. `flowMap` implements a nonparametric variate test to compare cell populations between samples. Our method can accommdate the high-dimensional features of FCM data and also the sample variation in distributions. Details of the method is described in [1].

2 Overview of the algorithm

`flowMap` implements the nonparametric Friedman-Rafsky (FR) multivariate run test to compare and match cell popoulations between samples. p-values of the FR statistic are calculated for each population comparison. Populations are considered a match if the p-values are above a Bonferroni-corrected cutoff. Moreover, we employed two approaches to calculate the p-values: 1) finding the percentile of the observed FR statistics in a standard normal distribution, and 2) finding the percentile of the statistics in a empirical null distribution of the FR statistics. The former assumes that the FR statistics follows a standard normal distribution, while the latter takes no distributional assumptions. Hereafter, the former p-value is referred to as the *theoretical p-value*, and the latter is referred to as the *empirical p-value*. Both p-values are provided in the output.

The goal of our algorithm is to identify matched versus mismatched cell populations by comparing each FCM sample to a selected rerference sample. Thus, two cell populations matched to the same reference are considered to be similar to each other as well. This method reduces the computational complexity. In addition, the algorithm allows the user to choose a reference sample for mapping or to construct a reference sample from the FCM test samples. The general flow of the diagram is as follows:
Denote S_o as the reference sample with m populations, and S_i as the test sample with n_i populations where $i = 1, \ldots, n$. Then,

1. Compare every S_i with S_o,

   Step 1: compute FR statistics of the $n_i \times m$ population pairs,

Step 2: compute p-values for the FR statistics,

Step 3: identify a population pair as matched if the p-value is less than $0.01/(m \times n_i)$,

2. Reassign S_i population labels to the matched S_o population label or a new unique population lable if there is no match in S_o.
3. Make a metaset of cell population labels by combining the matched and mismatched populations across all samples $S_1, \ldots, S_n$.

3 Data preparation

Here we assume that the data have been normalized and transformed according to appropriate flow cytometry data procedure. The input data can be in txt format or as data.frames, where the rows are the event (cell) data. The columns are consisted of the features and also the cell population identifying number as the last column of the data. Below is an example data `Sample1`. There are 25,809 events in total with 5 feature markers (CD20,CD24,CD27,CD38,IgD). The last column of the data indexes the cell population labels.

```
> sam1 <- read.table(system.file("extdata/sample1.txt"
+             ,package="flowMap"),header=T)
> str(sam1)
```

```
'data.frame':        13531 obs. of  6 variables:
 $ CD20: int  1728 1226 977 1322 3184 3064 1339 1356 1343 3211 ...
 $ CD24: int  517 477 418 540 2126 1995 645 373 337 683 ...
 $ CD27: int  30 753 499 1605 799 1 490 821 522 30 ...
 $ CD38: int  2795 2539 2541 2499 2076 2038 2325 3115 2415 1964 ...
 $ IgD : int  1396 1652 1435 1509 1918 1224 1500 1456 1508 1506 ...
 $ id  : int  1 1 1 1 2 2 1 1 1 2 ...
```

4 Mapping cell populations across two samples

For a two-sample comparison of $m \times n$ population pairs, we estimate the FR statistics for each pair with median FR statistics across D random draws. Each random draw samples k events from each sample, with a total of $2k$ events per random draw in a population comparison. The empirical null distribution of the FR statistic is calculated by shuffling cell population labels across two-samples. `flowMap` provides a comphrehensive parameter setting. Finally, we build in `doParallel` for user to specify the number of processing cores for the mapping

function. In the case when the number of requested processing cores is greater than the number of processing cores available, the function will opt for the maximum number avaiable cores in the system.

In the following example, median FR statistic for each population pair is estimated from 5 random draws (`draws=5`), with 100 events per population comparison (`sampleSize=100`). The cutoff for the p-values is set at 0.01/30 (`cutoff=`$10^{(}-5$)). The number of permutation for building empirical null distribution is set at 300 (`nperm=300`). 10 processing cores are requested.

The 2 populations in `Sample 1` are compared with the 1 population in `Sample 2`.

```
> sam1 <- read.table(system.file("extdata/sample1.txt"
+                 ,package="flowMap"),header=T)
> sam2 <- read.table(system.file("extdata/sample2.txt"
+                 ,package="flowMap"),header=T)
> table(sam1$id)
```

```
   1    2
9821 3710
```

```
> table(sam2$id)
```

```
   1
9821
```

The `FRmappingSimple` function computes FR statistic, p-values, and generates a list of matched/mismatched population labels.

```
> require(flowMap)
> res1 <- FRmappingSimple(samples=list(sam1),centroids=NULL,
+           refSample=sam2,refCentroid=NULL,nPopFilt=NULL,draws=5,cutoff=10^(-5),
+           sampleMethod="equalSize",sampleSize=100,nperm=300)
```

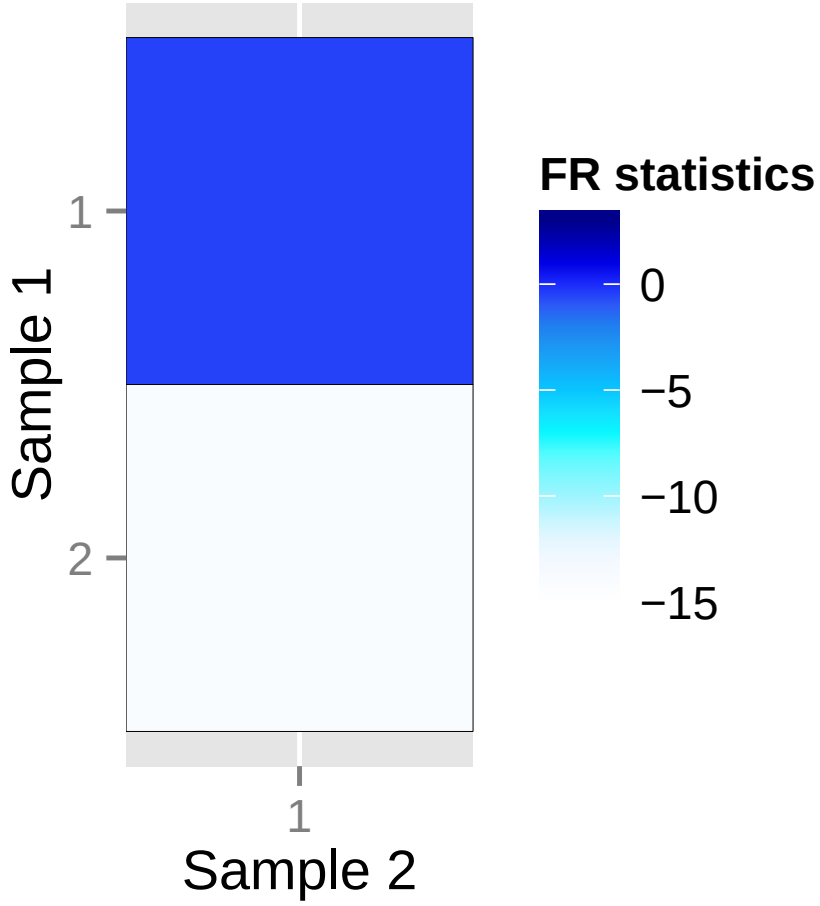
Return the estimated FR statistics

```
> str(getFRmapStats(res1))
```

```
List of 1
 $ : num [1:2, 1] -0.426 -14.069
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "1" "2"
  .. ..$ : chr "1"
```

The estimated FR statistics is -0.426 for CP1 in Sample 1 compared with CP1 in Sample 2, and the estimated FR statistics is -14.068 for CP2 in Sample 1 compared with CP1 in Sample 2.

```
> datToPlot <- melt(getFRmapStats(res1)[[1]])
> colnames(datToPlot) <- c("X1","X2","value")
> datToPlot$value[is.na(datToPlot$value)] <- -200
> pp <- ggplot(datToPlot, aes(factor(X2), factor(X1,levels=rev(sort(unique(datToPlot$X1))) ) )
+   coord_fixed() + geom_tile(aes(fill = value),colour="black") +
+    scale_fill_gradientn("FR statistics",colours=c("darkblue","blue","dodgerblue2",
+
+                                                                    values=rescale(c(2,1,0,
+                                                                    guide="colorbar",limits
+    labs(x="Sample 2",y="Sample 1")
> print(pp)
```



Below are results of theoretical and empirical p-values. Both p-values suggest that between the two cell populations in Sample 1, CP1 in Sample 1 is more similar to CP1 in Sample 2 than CP2 in Sample 1.

```
> str(getFRmapPnorms(res1))
```

```
List of 1
 $ : num [1:2, 1] 3.35e-01 2.95e-45
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "1" "2"
  .. ..$ : chr "1"
```

```
> str(getFRmapPnulls(res1))
```

```
List of 1
 $ : num [1:2, 1] 9.37e-01 1.00e-45
```

```
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "1" "2"
  .. ..$ : chr "1"
```

Below return the matching results of cell populations. CP1 in Sample 1 is mapped to CP1 in Sample 2, and CP2 in Sample 1 is mapped to CP1 in Sample 2.

```
> getCrossList(res1)
```

```
  sampleID testSample refSample newID
1        1          1         1     1
2        1          2        NA     2
```

5 References

[1] Chiaowen Hsiao, Mengya Liu, Yu Qian, Monnie McGee, and Richard Scheuermann (2013). Multiple sample comparison in flow cytometry data (manuscript in preparation).

6 SessionInfo

```
> toLatex(sessionInfo())
```

- R version 3.0.2 (2013-09-25), i386-w64-mingw32
- Locale: LC_COLLATE=C, LC_CTYPE=English_United States.1252, LC_MONETARY=English_United States.1252, LC_NUMERIC=C, LC_TIME=English_United States.1252
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: abind 1.4-0, ade4 1.5-2, doParallel 1.0.3, flowMap 1.0.0, foreach 1.4.1, ggplot2 0.9.3.1, iterators 1.0.6, reshape2 1.2.2, scales 0.2.3
- Loaded via a namespace (and not attached): MASS 7.3-29, RColorBrewer 1.0-5, codetools 0.2-8, colorspace 1.2-4, compiler 3.0.2, dichromat 2.0-0, digest 0.6.3, grid 3.0.2, gtable 0.1.2, labeling 0.2, munsell 0.4.2, plyr 1.8, proto 0.3-10, stringr 0.6.2, tools 3.0.2

```
list()
```