1 Introduction

The Gene Set BenchMark contains eleven expression datasets representative of different diseases. The package also contains a list of pathways and their associated gene targets. Together with these datasets and the pathways provide a benchmark for machine learning and pathway analyses, most of them used previously in [1].

2 Datasets

Benchmark datasets and pathway targets were downloaded from supplemental files and sources cited in [1]. These datasets covers different diseases: Leukaemia [2], Marfan [3],
Melanoma [4], Prostate [5], Sarcoma [6], Head and neck cancer [7], response to breast cancer treatment [8], Bipolar disorder [9]. We also added datasets for two new diseases: Parkinson’s disease [10], and Melanoma cancer [4]. We did not include two of the datasets mentioned in [1]: First, the famous Leukemia data set in [11] which is available through package golubEsets. Secondly, the data presented in paper [12] because the data were not available to us. These data were converted from Matlab to R for this package.

First, we load the library:

```r
> require(GSBenchMark)
```

## 2.1 Pathway Data

**GSBenchMark** contains a list of the pathways.

```r
> data(diracpathways)
> class(diracpathways)
[1] "list"
> names(diracpathways)[1:5]
[1] "DEATHPATHWAY" "TCAPOPTOSISPATHWAY" "CCR3PATHWAY" "NEUTROPHILPATHWAY" "ALTERNATIVEPATHWAY"
> class(diracpathways[[1]])
[1] "character"
> diracpathways[[1]]

"BID" "TRAF2" "TNFRSF25" "NFKBIA" "NFKB1" "TNFSF12" "CASP6" "CASP3" "CASP9" "CASP7" "BCL2" "CASP8" "CHUK" "CFLAR" "DFFA" "DFFB" "RELA" "CYCS" "LMNA" "GAS2" "FADD" "BIRC4" "BIRC3" "BIRC2" "TRADD" "TNFRSF10A" "CASP10" "TNFSF10" "TNFRSF10B" "RIPK1" "APAF1" "MAP3K14" "SPTAN1"
> pathways = diracpathways;
```

The variable `diracpathways` contains the pathways genes. It is a list. Each element represents a pathway with its name. Each elements contains a list of characters which represent the genes in the pathway.
2.2 Gene Expression Datasets

Now, we load the datasets names:

```r
> data(GSBenchMarkDatasets)
> print(GSBenchMark.Dataset.names)

[1] "leukemia_GSEA"    "marfan_GDS2960"    "melanoma_GDS2735"
[4] "parkinsons_GDS2519" "prostate_GDS2545_m_nf" "prostate_GDS2545_m_p"
[7] "prostate_GDS2545_p_nf" "sarcoma_data"    "squamous_GDS2520"
[10] "breast_GDS807"      "bipolar_GDS2190"
```

Here is a summary of the datasets:

```r
> for(i in 1: length(GSBenchMark.Dataset.names))
> {
>   DataSetStudy = GSBenchMark.Dataset.names[[i]]
>   data(list=DataSetStudy)
>   cat("Data Set",DataSetStudy,\"n\")
>   print(phenotypesLevels)
>   print(table(phenotypes))
> }

Data Set leukemia_GSEA
   0 1
  "AML" "ALL"
phenotypes
   0 1
  24 24

Data Set marfan_GDS2960
   0 1
  "non-MFS" "MFS"
phenotypes
   0 1
  41 60

Data Set melanoma_GDS2735
   0 1
  "Normal" "metastasis"
phenotypes
   0 1
  23 23

Data Set parkinsons_GDS2519
   0 1
  "Normal" "Parkinson's"
phenotypes
   0 1
  22 50

Data Set prostate_GDS2545_m_nf
   0 1
  "normal" "metastasis"
phenotypes
   0 1
```
The data consist of three variables: \texttt{exprsdata}, \texttt{phenotypes}, and \texttt{phenotypesLevels}. \texttt{exprsdata} consists of gene expressions. \texttt{phenotypes} contains the sample labels: "$0$" indicates less dangerous and "$1$" more dangerous phenotype. \texttt{phenotypesLevels} makes the connection between "$0$" and "$1$" with the real label names e.g. "Normal" and "Parkinson's". \texttt{GSBenchMark} requires the rownames of gene expression varaible represent the gene names, \textit{i.e.} they are represented in the pathway information variable.

### 2.3 Matching pathway targets to gene expression datasets

One can extract the gene names by:
Also, it is possible that some of the genes in a pathway are not represented in the expression data. We prune them as the following:

```r
> prunedpathways = lapply(pathways, function(x) intersect(x, genenames))
> emptypathways = which(sapply(prunedpathways, length) < 2)
> if (length(emptypathways) > 0) {
    warning(paste("After pruning the pathways, there exist pathways with zero or one gene!\nSmall pathways were deleted. Deleted pathways: ", paste(names(emptypathways), collapse = ","), "\n"))
    diracpathwayPruned= prunedpathways[-emptypathways]
} else {
    diracpathwayPruned = prunedpathways
}
> cat("Number of pathways before pruning ",length(pathways),"\n")
Number of pathways before pruning  249
> cat("Number of pathways after pruning ",length(diracpathwayPruned))
Number of pathways after pruning  249
```

**phenotypes** is a factor with with levels ("0","1") where "1" indicates more dangerous phenotype. For real labels, one can look at **phenotypesLevels**

```r
> summary(phenotypes)
0 1
31 30

> phenotypesLevels

0 1
"Normal" "Bipolar"
```
3 System Information

Session information:

\texttt{\textgreater{} toLatex(sessionInfo())}

- R version 3.3.1 (2016-06-21), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8,
  LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8,
  LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C,
  LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: GSBenchMark 0.108.0
- Loaded via a namespace (and not attached): tools 3.3.1

4 Literature Cited

References


