Gene Set BenchMark

Bahman Afsari[1] and Elana J. Fertig[1]

[1]The Sidney Kimmel Comprehensive Cancer Center,
Johns Hopkins University School of Medicine

Modified: April 8, 2014. Compiled: November 5, 2019

Contents

1   Introduction

The `GSBenchMark` 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],

1

Melanoma [4], Prostate [5], Sarcoma [6], Head and neck cancer [7], response to breast cancer treatmetn [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 Leaukemia 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:

```
> require(GSBenchMark)
```

2.1 Pathway Data

GSBenchMark contains a list of the pathways.

```
> data(diracpathways)
> class(diracpathways)

[1] "list"

> names(diracpathways)[1:5]

[1] "DEATHPATHWAY"      "TCAPOPTOSISPATHWAY" "CCR3PATHWAY"
[4] "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:

```
> 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:

```
> 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
```

```
18 25
Data Set prostate_GDS2545_m_p
             0             1
   "primary" "metastasis"
phenotypes
 0  1
65 25
Data Set prostate_GDS2545_p_nf
         0           1
 "normal" "primary"
phenotypes
 0  1
18 65
Data Set sarcoma_data
       0       1
 "LMS" "GIST"
phenotypes
 0  1
31 37
Data Set squamous_GDS2520
         0         1
"Normal"   "HNSCC"
phenotypes
 0  1
22 22
Data Set breast_GDS807
                 0                 1
    "Responsive" "Non-responsive"
phenotypes
 0  1
32 28
Data Set bipolar_GDS2190
         0         1
 "Normal" "Bipolar"
phenotypes
 0  1
31 30
```

The data consist of three variables: `exprsdata`, `phenotypes`, and `phenotypesLevels`. `exprsdata` consists of gene expressions. `phenotypes` contains the sample labels: "0" indicates less dangarous and "1" more dangerous phenotype. `phenotypesLevels` makes the connection between "0" and "1" with the real label names e.g. "Normal" and "Parkinson's". `GSBenchMark` requires the rownames of gene expression varaible represent the gene names, *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:

```
> genenames = rownames(exprsdata);
> genenames[1:10]

 [1] "DDR1"    "RFC2"    "HSPA6"   "PAX8"    "GUCA1A" "UBA7"    "THRA"    "PTPN21" "CCL5"
[10] "CYP2E1"
```

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:

```
> 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!\n Sma
            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**

```
> summary(phenotypes)

 0  1
31 30

> phenotypesLevels

        0        1
  "Normal" "Bipolar"
```

3   System Information

Session information:

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

- R version 3.6.1 (2019-07-05), 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

- Running under: Ubuntu 18.04.3 LTS

- Matrix products: default

- BLAS: /home/biocbuild/bbs-3.10-bioc/R/lib/libRblas.so

- LAPACK: /home/biocbuild/bbs-3.10-bioc/R/lib/libRlapack.so

- Base packages: base, datasets, grDevices, graphics, methods, stats, utils

- Other packages: GSBenchMark 1.6.0

- Loaded via a namespace (and not attached): compiler 3.6.1, tools 3.6.1

4   Literature Cited

References

[1] James A Eddy, Leroy Hood, Nathan D Price, and Donald Geman. Identifying tightly
regulated and variably expressed networks by differential rank conservation (dirac).
PLoS computational biology, 6(5):e1000792, 2010.

[2] Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L
den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and
Stanley J Korsmeyer. Mll translocations specify a distinct gene expression profile that
distinguishes a unique leukemia. *Nature genetics*, 30(1):41–47, 2002.

[3] Zizhen Yao, Jochen C Jaeger, Walter L Ruzzo, Cecile Z Morale, Mary Emond, Uta
Francke, Dianna M Milewicz, Stephen M Schwartz, and Eileen R Mulvihill. A mar-
fan syndrome gene expression phenotype in cultured skin fibroblasts. *BMC genomics*,
8(1):319, 2007.

[4] Rebecca J Critchley-Thorne, Ning Yan, Serban Nacu, Jeffrey Weber, Susan P Holmes, and Peter P Lee. Down-regulation of the interferon signaling pathway in t lymphocytes from patients with metastatic melanoma. *PLoS Medicine*, 4(5):e176, 2007.

[5] Uma R Chandran, Changqing Ma, Rajiv Dhir, Michelle Bisceglia, Maureen Lyons-Weiler, Wenjing Liang, George Michalopoulos, Michael Becich, and Federico A Monzon. Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC cancer*, 7(1):64, 2007.

[6] Nathan D Price, Jonathan Trent, Adel K El-Naggar, David Cogdell, Ellen Taylor, Kelly K Hunt, Raphael E Pollock, Leroy Hood, Ilya Shmulevich, and Wei Zhang. Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. *Proceedings of the National Academy of Sciences*, 104(9):3414–3419, 2007.

[7] MA Kuriakose, WT Chen, ZM He, AG Sikora, P Zhang, ZY Zhang, WL Qiu, DF Hsu, C McMunn-Coffran, SM Brown, et al. Selection and validation of differentially expressed genes in head and neck cancer. *Cellular and Molecular Life Sciences CMLS*, 61(11):1372–1383, 2004.

[8] Xiao-Jun Ma, Zuncai Wang, Paula D Ryan, Steven J Isakoff, Anne Barmettler, Andrew Fuller, Beth Muir, Gayatry Mohapatra, Ranelle Salunga, J Todd Tuggle, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer cell*, 5(6):607–616, 2004.

[9] MM Ryan, HE Lockstone, SJ Huffaker, MT Wayland, MJ Webster, and S Bahn. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Molecular psychiatry*, 11(10):965–978, 2006.

[10] Clemens R Scherzer, Aron C Eklund, Lee J Morse, Zhixiang Liao, Joseph J Locascio, Daniel Fefer, Michael A Schwarzschild, Michael G Schlossmacher, Michael A Hauser, Jeffery M Vance, et al. Molecular markers of early parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences*, 104(3):955–960, 2007.

[11] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[12] Carlos S Moreno, Lilya Matyunina, Erin B Dickerson, Nina Schubert, Nathan J Bowen, Sanjay Logani, Benedict B Benigno, and John F McDonald. Evidence that p53-mediated cell-cycle-arrest inhibits chemotherapeutic treatment of ovarian carcinomas. *PLoS One*, 2(5):e441, 2007.