sambar_tutorial

August 5, 2020

1 SAMBAR usage example

1.0.1 Author:

Genís Calderer*.
*Kuijjer Lab (NCMM) - genis.calderer@gmail.com

1.1 Introduction

SAMBAR is a method to identify subtypes of samples based on annotated mutation data. It uses gene mutation counts and membership of those genes in a set of pathways to summarize the mutation rate in a "pathway mutation score" thus despasifying the mutation data. For each sample and each pathway a score is computed and it can be used to compare the samples to identify subtypes. The SAMBAR package uses binomial distance and hierarchical clustering to find subgroups of samples with similar pathway mutation patterns. The method was developed and published in the following paper: Cancer subtype identification using somatic mutation data, Kuijjer ML, Paulson JN, Salzman P, Ding W, Quackenbush J, British Journal of Cancer (May 16, 2018), doi: 10.1038/s41416-018-0109-7, https://www.nature.com/articles/s41416-018-0109-7, BioRxiv, doi: https://doi.org/10.1101/228031

SAMBAR, or Subtyping Agglomerated Mutations By Annotation Relations, is a method to identify subtypes based on somatic mutation data. SAMBAR was used to identify mutational subtypes in 23 cancer types from The Cancer Genome Atlas (Kuijjer ML, Paulson JN, Salzman P, Ding W, Quackenbush J, *British Journal of Cancer* (May 16, 2018), doi: 10.1038/s41416-018-0109-7, https://www.nature.com/articles/s41416-018-0109-7, *BioRxiv*, doi: https://doi.org/10.1101/228031).

SAMBAR's input is a matrix that includes the number of non-synonymous mutations in a sample i and gene j. SAMBAR first subsets these data to a set of 2,219 cancer-associated genes (optional) from the Catalogue Of Somatic Mutations In Cancer (COSMIC) and Östlund *et al*. (Network-based identification of novel cancer genes, 2010, *Mol Cell Prot*), or from a user-defined list. It then divides the number of non-synonymous mutations by the gene's length L_j, defined as the number of non-overlapping exonic base pairs of a gene. For each sample, SAMBAR then calculates the overall cancer-associated mutation rate by summing mutation scores in all cancer-associated genes j'. It removes samples for which the mutation rate is zero and divides the mutation scores the remaining samples by the sample's mutation rate, resulting in a matrix of mutation rate-adjusted scores G:

$$G_{ij} = \frac{N_{ij}/L_j}{\sum_{j'}(N_{ij'}/L_{j'})}.$$

The next step in SAMBAR is de-sparsification of these gene mutation scores (agglomerated mutations) into pathway mutation (annotation relation) scores. SAMBAR converts a (user-defined) gene signature (.gmt format) into a binary matrix M, with information of whether a gene j belongs to a pathway q. It then calculates pathway mutation scores P by correcting the sum of mutation scores of all genes in a pathway for the number of pathways q' a gene belongs to, and for the number of cancer-associated genes present in that pathway:

$$P_{iq} = \frac{\sum_{j \in q} G_{ij} / \sum_{q'} M_{jq'}}{\sum_{j} M_{jq}}$$

Finally, SAMBAR uses binomial distance to cluster the pathway mutation scores. The cluster dendrogram is then divided into k groups (or a range of k groups), and the cluster assignments are returned in a list.

This guide will use the toy data included in the SAMBAR package to showcase the usage of this package.

1.2 1. Importing SAMBAR from netZooPy

In order to use the SAMBAR functions it has to be imported from the netZooPy as follows:

[11]:
```
from netZooPy import sambar
```

To see the parameters of the main function one can use the following line:

[ ]:
```
help(sambar.sambar)
```

1.3 2. Selecting input files

The program requires a gene mutation dataset, a list of gene sizes, a list of cancer-associated genes (optional) and a list of pathways with its genes.

[ ]:
```
# These are the names of the files of the toy dataset.
# The program by default runs with the toy data.
mut_file = "../tests/sambar/ToyData/mut.ucec.csv"
cangenes = "../tests/sambar/ToyData/genes.txt"
sign_file = "h.all.v6.1.symbols.gmt"
esize_file = "esizef.csv"
```

1.4 3. Run SAMBAR

The main SAMBAR function takes as input the filepaths of the datasets and returns a pathway score dataframe and a sample clustering dataframe for different cuts in the linkage tree. It also outputs a csv file for the adjusted mutation scores, pathway scores and clustering. The slow step in this method is the computation of the distance matrix, this matrix is also exported in case it's

needed and rerunning the whole process is not wanted. The function runs first the desparcification and then the clustering.

[ ]:
```
pathway_scores, cluster_groups = sambar.
↪sambar(mut_file,esize_file,cangenes,sign_file)
```

[15]:
```
pathway_scores, cluster_groups = sambar.sambar() #Runs with the default files.
```

```
Sambar runtime:  2.9713971614837646
Clustering runtime:  4.223911762237549
```

1.5 4. Results

The pathway mutation scores and the sample groups are the output of the method.

[16]:
```
pathway_scores.head(10)
```

[16]:

	TCGA-A5-A0G3	TCGA-A5-A0G5	TCGA-A5-A0G9	\
HALLMARK_ADIPOGENESIS	0.000000	0.000000	0.001195	
HALLMARK_ALLOGRAFT_REJECTION	0.002480	0.000000	0.000266	
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000000	0.001152	
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.001392	
HALLMARK_APICAL_JUNCTION	0.000000	0.000000	0.001352	
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.000000	
HALLMARK_APOPTOSIS	0.000875	0.000000	0.000311	
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000987	0.000000	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000000	0.000000	0.000000	
HALLMARK_COAGULATION	0.001983	0.000000	0.000000	

	TCGA-A5-A0GA	TCGA-A5-A0GB	TCGA-A5-A0GD	\
HALLMARK_ADIPOGENESIS	0.000369	0.000409	0.0	
HALLMARK_ALLOGRAFT_REJECTION	0.000000	0.000124	0.0	
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000000	0.0	
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.0	
HALLMARK_APICAL_JUNCTION	0.000364	0.000152	0.0	
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.0	
HALLMARK_APOPTOSIS	0.000238	0.000085	0.0	
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000000	0.0	
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.000814	0.000105	0.0	
HALLMARK_COAGULATION	0.000633	0.000102	0.0	

	TCGA-A5-A0GE	TCGA-A5-A0GH	TCGA-A5-A0GJ	\
HALLMARK_ADIPOGENESIS	0.000000	0.000000	0.000000	
HALLMARK_ALLOGRAFT_REJECTION	0.000000	0.000000	0.000000	
HALLMARK_ANDROGEN_RESPONSE	0.000000	0.000000	0.000768	
HALLMARK_ANGIOGENESIS	0.000000	0.000000	0.000000	
HALLMARK_APICAL_JUNCTION	0.005127	0.000167	0.000000	
HALLMARK_APICAL_SURFACE	0.000000	0.000000	0.000000	
HALLMARK_APOPTOSIS	0.000000	0.000000	0.000000	
HALLMARK_BILE_ACID_METABOLISM	0.000000	0.000000	0.000000	

```
HALLMARK_CHOLESTEROL_HOMEOSTASIS      0.000000      0.000000      0.000000
HALLMARK_COAGULATION                  0.000000      0.000000      0.000000

                                  TCGA-A5-A0GM  ...  TCGA-D1-A160  \
HALLMARK_ADIPOGENESIS                 0.000000  ...      0.000525
HALLMARK_ALLOGRAFT_REJECTION          0.000000  ...      0.000381
HALLMARK_ANDROGEN_RESPONSE            0.000000  ...      0.000395
HALLMARK_ANGIOGENESIS                 0.000000  ...      0.000000
HALLMARK_APICAL_JUNCTION              0.000703  ...      0.000313
HALLMARK_APICAL_SURFACE               0.000000  ...      0.000000
HALLMARK_APOPTOSIS                    0.001057  ...      0.000084
HALLMARK_BILE_ACID_METABOLISM         0.000000  ...      0.000000
HALLMARK_CHOLESTEROL_HOMEOSTASIS      0.003094  ...      0.000000
HALLMARK_COAGULATION                  0.000000  ...      0.000253

                                  TCGA-D1-A161  TCGA-D1-A167  TCGA-D1-A16F  \
HALLMARK_ADIPOGENESIS                 0.000000      0.000507      0.000000
HALLMARK_ALLOGRAFT_REJECTION          0.000645      0.000161      0.000111
HALLMARK_ANDROGEN_RESPONSE            0.000000      0.000231      0.000000
HALLMARK_ANGIOGENESIS                 0.000000      0.000000      0.000000
HALLMARK_APICAL_JUNCTION              0.000431      0.000429      0.000278
HALLMARK_APICAL_SURFACE               0.000000      0.000000      0.000000
HALLMARK_APOPTOSIS                    0.000853      0.000184      0.000220
HALLMARK_BILE_ACID_METABOLISM         0.000000      0.000217      0.000000
HALLMARK_CHOLESTEROL_HOMEOSTASIS      0.000000      0.000000      0.000000
HALLMARK_COAGULATION                  0.000000      0.000757      0.002150

                                  TCGA-D1-A16X  TCGA-D1-A16Y  TCGA-D1-A17Q  \
HALLMARK_ADIPOGENESIS                 0.000102      0.000034      0.000266
HALLMARK_ALLOGRAFT_REJECTION          0.000210      0.000146      0.000257
HALLMARK_ANDROGEN_RESPONSE            0.000040      0.000000      0.000182
HALLMARK_ANGIOGENESIS                 0.000000      0.000000      0.000424
HALLMARK_APICAL_JUNCTION              0.000358      0.000325      0.000245
HALLMARK_APICAL_SURFACE               0.000000      0.000000      0.000088
HALLMARK_APOPTOSIS                    0.000248      0.000400      0.000258
HALLMARK_BILE_ACID_METABOLISM         0.000306      0.000000      0.000299
HALLMARK_CHOLESTEROL_HOMEOSTASIS      0.000000      0.000000      0.000028
HALLMARK_COAGULATION                  0.000235      0.001148      0.000455

                                  TCGA-D1-A1NX  TCGA-EY-A1GS  TCGA-EY-A212
HALLMARK_ADIPOGENESIS                 0.000000      0.000210           0.0
HALLMARK_ALLOGRAFT_REJECTION          0.000000      0.000000           0.0
HALLMARK_ANDROGEN_RESPONSE            0.000000      0.000000           0.0
HALLMARK_ANGIOGENESIS                 0.000000      0.000000           0.0
HALLMARK_APICAL_JUNCTION              0.000519      0.000085           0.0
HALLMARK_APICAL_SURFACE               0.000000      0.000000           0.0
HALLMARK_APOPTOSIS                    0.000000      0.000211           0.0
```

```
HALLMARK_BILE_ACID_METABOLISM          0.000000      0.000000          0.0
HALLMARK_CHOLESTEROL_HOMEOSTASIS       0.000000      0.000000          0.0
HALLMARK_COAGULATION                   0.001008      0.001918          0.0

[10 rows x 247 columns]
```

[17]:
```
cluster_groups.head()
```

[17]:
```
    TCGA-A5-A0G3  TCGA-A5-A0G5  TCGA-A5-A0G9  TCGA-A5-A0GA  TCGA-A5-A0GB  \
X2             0             0             0             1             1
X3             0             0             1             2             2
X4             0             0             1             2             3

    TCGA-A5-A0GD  TCGA-A5-A0GE  TCGA-A5-A0GH  TCGA-A5-A0GJ  TCGA-A5-A0GM  ...  \
X2             0             0             0             0             0  ...
X3             0             0             1             0             0  ...
X4             0             0             1             0             0  ...

    TCGA-D1-A160  TCGA-D1-A161  TCGA-D1-A167  TCGA-D1-A16F  TCGA-D1-A16X  \
X2             1             0             1             0             1
X3             2             1             2             1             2
X4             2             1             3             1             3

    TCGA-D1-A16Y  TCGA-D1-A17Q  TCGA-D1-A1NX  TCGA-EY-A1GS  TCGA-EY-A212
X2             1             1             0             1             0
X3             2             2             0             2             0
X4             3             3             0             2             0

[3 rows x 247 columns]
```

[ ]: