The SVM2CRM data User's Guide

Guidantonio Malagoli Tagliazucchi, Silvio Bicciato
Center for genome research
University of Modena and Reggio Emilia, Modena

May 20, 2021

May 20, 2021

Contents

1 Overview

The SVM2CRMdata package contains ChIP-seq data of methylation and acetylation maps from CD4+ T-cells([1]). The package contains also a list of p300 binding sites and two matrix contains the signals of the histone marks in correspondance of genomic regions with enhancers or background sequences. This package is useful to perform the tutorial analysis of SVM2CRM. In this tutorial we load a matrix with the signal of ChIP-seq data preprocess using cisREfindbed function of SVM2CRM package and two matrices containing the signal of the histone marks at level of putative enhancers and random genomic regions.

2 CD4matrixInputSVMbin100window1000 object

This data.frame contains the signals of the histone modification from CD4+T-cells([1]) genome wide. In particular, this dataset contains 20 methylation and 17 acetylation ChIP-seq maps. We obtained this object using the function cisREfindbed from *SVM2CRM* package. For details about this function see the documentation in *SVM2CRM* package. Briefly,this function allowed the user to create one of the input require to perfomed the prediction of enhancers using SVM2CRM. In particular, this function reads external bed files of the histone modification (processed externally from R) and performed the smoothing of the signals genome-wide. When we processed this dataset we used cisREfindbed considering a bin.size of 100 bp, that is the values that we used to binarized the signals of the histone modification genome wide. Moreover the size of the window to perfomed the smoothing of the histone modification signals genome wide was set to 500, 1000, 2000 bp. This package contains the data.frame obtained using a window of 1000bp and considering only the chromosome 1. Before to perfomed the prediction of enhancers using your dataset and SVM2CRM we suggests to generate this data.frame using ad hoc script that contain only the code of cisREfindbed and that save the results of this analysis in a .RData. This can help the user to reduce the computational time to perfomed the prediction genome-wide and avoid to create the input each time that the user want launch a new analysis (for e.g. using different parameters).

```
> library(SVM2CRMdata)
> setwd(system.file("data",package="SVM2CRMdata"))
> load("CD4_matrixInputSVMbin100window1000.rda")
> #The column contains the signal. The rows contains the genomic coordinate, while th
> dim("CD4_matrixInputSVMbin100window1000")

NULL
```

3 trainpositive and trainnegative objects

The trainpositive and trainnegative objects were created using the function getSignal of *SVM2CRM* package. These data.frame contains the signals of the histone modification of 20 methylation and 17 acetylation ChIP-seq maps from ([1]) at p300 binding sites (train positive) and genomic random regions (train negative). In particular these data.frame contains the signals of the histone modifications in correspondence of p300 binding sites known

and genomic regions distant from TSSs of 1000bp. The creation of this matrices is mandatory because SVM2CRM require the presence of two datasets that contains the signals of the histone modification at level of putative enhancers (p300 binding sites) and in genomic regions different from distal cis-regulatory regions. Here we used getSignal from *SVM2CRM* to create respectivelly the trainpositive and trainnegative dataset. In particular, we download a list p300 binding sites and create a list of genomic-coordinates random regions using bedtools using a windows size of 1000. We used the same size of window (1000 bp) defined also to create the object described in the previously object (CD4matrixInputSVMbin100window1000).getSignal consider the signal of the histone marks around a particular windows (1000 bp) from the start of the genomic regions provided (e.g. the start of the p300 binding sites and background genomic regions). Also in this case we suggests to create this data.frame before to perfomed an entire analysis of SVM2CRM.

```
> setwd(system.file("data",package="SVM2CRMdata"))
> load("train_positive.rda")
> load("train_negative.rda")
```

4 External data

In *SVM2CRMdata* we provided ChIP-seq data from H2AK5ac, H2K23ac, H2AK9ac, H3K27ac, H3K4me1, H3K4me2, H3K4me3. This data contains the signals of the histone modification normalized using the procedure described in ([2]). The user can use these data to create the object described in the first section of this vignette. This object is one of the most important input required to perfomed the prediction of analysis with *SVM2CRM*. As described previously this object contains the genome wide signals of the histone modifications provided by the user. The user can create this input simple using cisREfindbed function from *SVM2CRM*. The parameters required are bin.size used to binnarized the signal of the histone modification (in this case 100bp), the size of the windows to consider the signals of the histone marks (each 1000 bp), and the function that the user want use to model the signals of the histone marks inside the windows (optional). The object that is generate using cisREfindbed several columns. A column with the chromosome, the start and end of the genomic windows, the signals of the histone marks inside the windows. The number of columns that contains the signals of the histone marks dependes on the window size. For details see section "Introduction" in the vignette of *SVM2CRM*

5 References

References

[1] Rosenfeld JA Wang Z, Zang C, Schones DE, Cuddapah S Barski A, Roh TY Cui K, Zhang MQ Peng W, and Zhao K. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 7(40):897–903, 2008.

[2] Hiram A Firpi, Duygu Ucar, and Kai Tan. Discover regulatory dna elements using chromatin signatures and artificial neural network. *Bioinformatics*, 26(13):1579–86, 2010.

6 Session information

The output in this vignette was produced under the following conditions:

```
> sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.13-bioc/R/lib/libRblas.so
LAPACK: /home/biocbuild/bbs-3.13-bioc/R/lib/libRlapack.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] SVM2CRMdata_1.24.0

loaded via a namespace (and not attached):
[1] compiler_4.1.0 tools_4.1.0
```