projectoR Vignette

Genevieve L. Stein-O'Brien

14 June 2017

Contents

1 Introduction

Technological advances continue to spur the exponential growth of biological data as illustrated by the rise of the omics—genomics, transcriptomics, epigenomics, proteomics, etc.—each with there own high throughput technologies. In order to leverage the full power of these resources, methods to integrate multiple data sets and data types must be developed. The reciprocal nature of the genomic, transcriptomic, epigenomic, and proteomic biology requires that the data provides a complementary view of cellular function and regulatory organization; however, the technical heterogeneity and massive size of high-throughput data even within a particular omic makes integrated analysis challenging. To address these challenges, we developed ProjectoR, an R package for integrated analysis of high dimensional omic data. ProjectoR uses the relationships defined within a given high dimensional data set, to interrogate related biological phenomena in an entirely new data set. By relying on relative comparisons within data type, ProjectoR is able to circumvent many issues arising from technological variation. For a more extensive example of how the tools in the ProjectorR package can be used for *in silico* experiments, or additional information on the algorithm, see Stein-O'Brien, et al.

2 Getting started with ProjectoR

2.1 Installation Instructions

For automatic Bioconductor package installation, start R, and run:

```
source("https://bioconductor.org/biocLite.R")
biocLite("projectoR")
```

For the current development level version, start R, and run:

```
library(devtools)
install_github("projectoR", "genesofeve")
```

2.2 Methods

Projection can roughly be defined as a mapping or transformation of points from one space to another often lower dimensional space. Mathematically, this can described as a function $\varphi(x) = y : \Re^D \mapsto \Re^d$ s.t. $d \leq D$ for $x \in \Re^D, y \in \Re^d$ [**?**]. The projectoR package uses projection functions defined in a training dataset to interrogate related biological phenomena in an entirely new data set. These functions can be the product of any one of several methods common to "omic" analyses including regression, PCA, NMF, clustering. Individual chapters focussing on one specific method are included in the vignette. However, the general design of the projectoR function is the same reguardless.

2.3 The base projectoR function

The base projectoR function is executed as follows:

```
library(projectoR)
projectoR(data = NA, AnnotionObj = NA, IDcol = "GeneSymbol",
    Patterns = NA, NP = NA, full = FALSE)
```

2.3.1 Input Arguments

The inputs that must be set each time are only the data and patterns, with all other inputs having default values. However, inconguities between gene names–rownames of the Patterns object and either rownames of the data object or the values in the "IDcol" of a corresponding annotation object for the data–will throw errors and, subsequently, should be checked before running.

The arguments are as follows:

data a dataset to be projected into the pattern space
AnnotionObj an annotion object for data. If NA the rownames of data will be used.
IDcol the column of AnnotionData object corresponding to identifiers matching the type used for GeneWeights
Patterns a matrix of continous values with unique rownames to be projected
NP vector of integers indicating which columns of Patterns object to use. The default of NP = NA will use entire matrix.
full logical indicating whether to return the full model solution. By default only the new pattern object is returned.

The `Patterns` arguement in the base projectoR function is suitable for use with any genernal feature space, or set of feature spaces, whose rows annotation links them to the data to be projected. Ex: the coeffients associated with individual genes as the result of regression analysis or the amplituded values of individual genes as the result of non-negative matrix factorization (NMF).

2.3.2 Output

The basic output of the base projectoR function, i.e. `full=FALSE`, returns `projectionPatterns` representing relative weights for the samples from the new data in this previously defined feature space, or set of feature spaces. The full output of the base projectoR function, i.e. `full=TRUE`, returns `projectionFit`, a list containing `projectionPatterns` and `Projection`. The `Projection` object contains additional information from the proceedure used to obtain teh `projectionPatterns`. For the the the base projectoR function, `Projection` is the full lmFit model from the .

3 PCA projection

Projection of principle components is achieved by multiplying a new data set by previously generated eigenvectors, or gene loadings. The `projectoR` function has S3 method for class `prcomp`.

3.1 Obtaining PCs to project.

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

3.2 Projecting prcomp objects

```
## [1] 93  9
```

4 CoGAPS projection

5 Clustering projection

6 Correlation based projection