Toolbox for analysis and prediction of protein and peptide variant effects

22100 - R for Bio Data Science

Group 3: Felix Pacheco, Jacob Kofoed, Begoña Bolos Sierra, Laura Sans Comerma

Spring 2020

Content

- Introduction
- Methods
  - Project overview
  - Data
- Results
- Discussion

Introduction

Prediction of protein-protein interactions (PPI) are a challenging task.

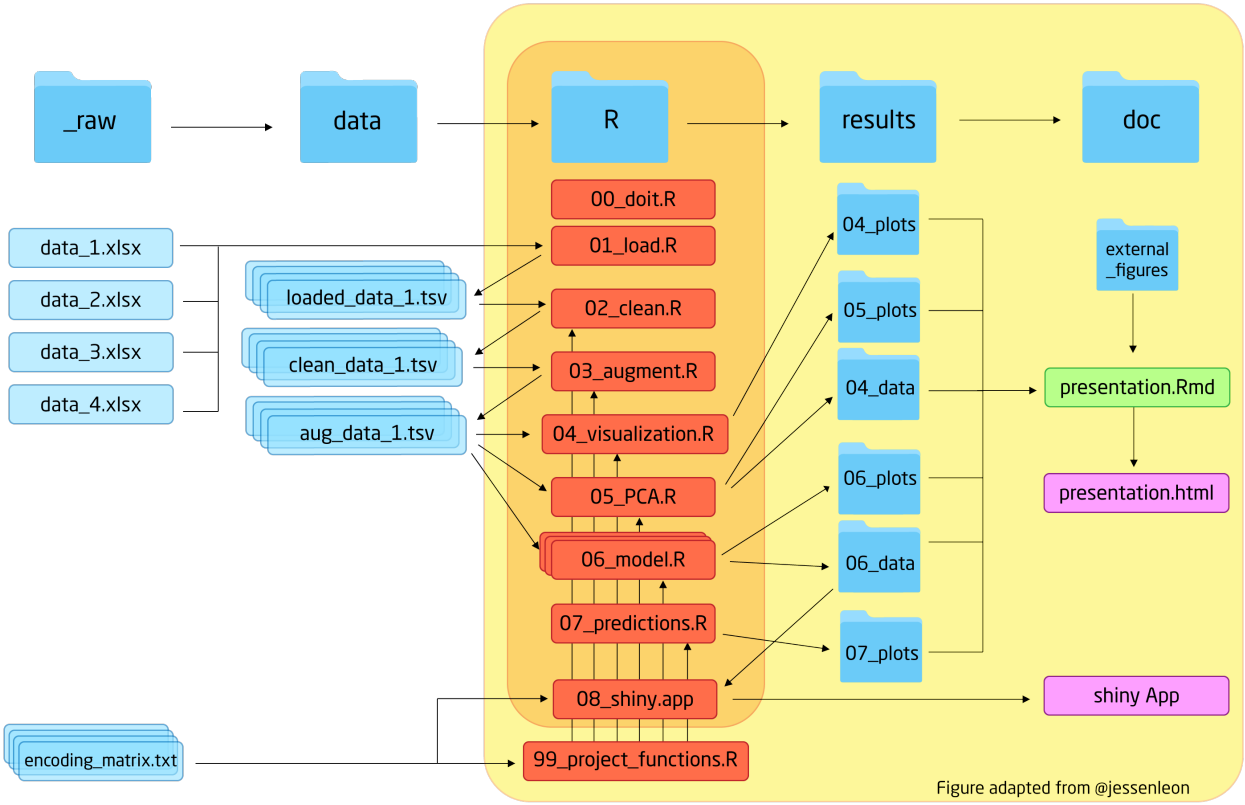
ML models allow to exploit the content of these PPI data sets.

The aim of this project is to create a toolbox to predict the biological activity of these peptides with machine learning models.

- Features:
  - Support for both sequence or variant input.
  - Support for several sequence encoders.
  - Support for several models.
  - Visualization options.

Project overview

Visit our Github repository



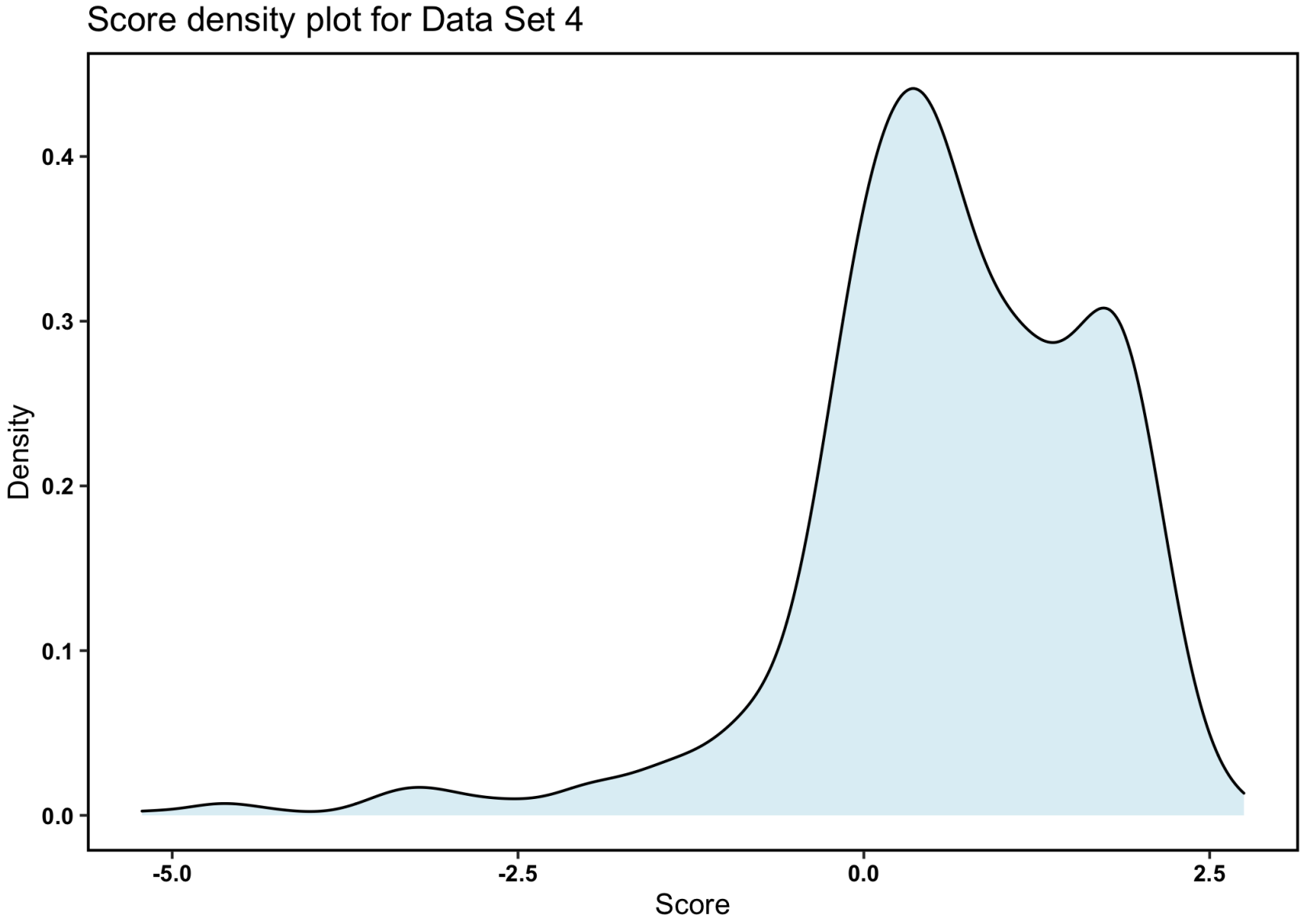
Methods - packages used

Function	Library
Data loading	`readxl`
Data cleaning and wrangling	`dplyr` , `broom` (`tidyverse`)
Data augmenting	`dplyr` (`tidyverse`),`Peptides`
Extracting data	`UniprotR`
Plotting	`ggplot2`(`tidyverse`), `ggseqlogo`,`ggpubr`
Analysing	`stats`
Modeling	`keras`,`neuralnet`, `caret`, `yardstick`, `gl`

Methods - the data sets

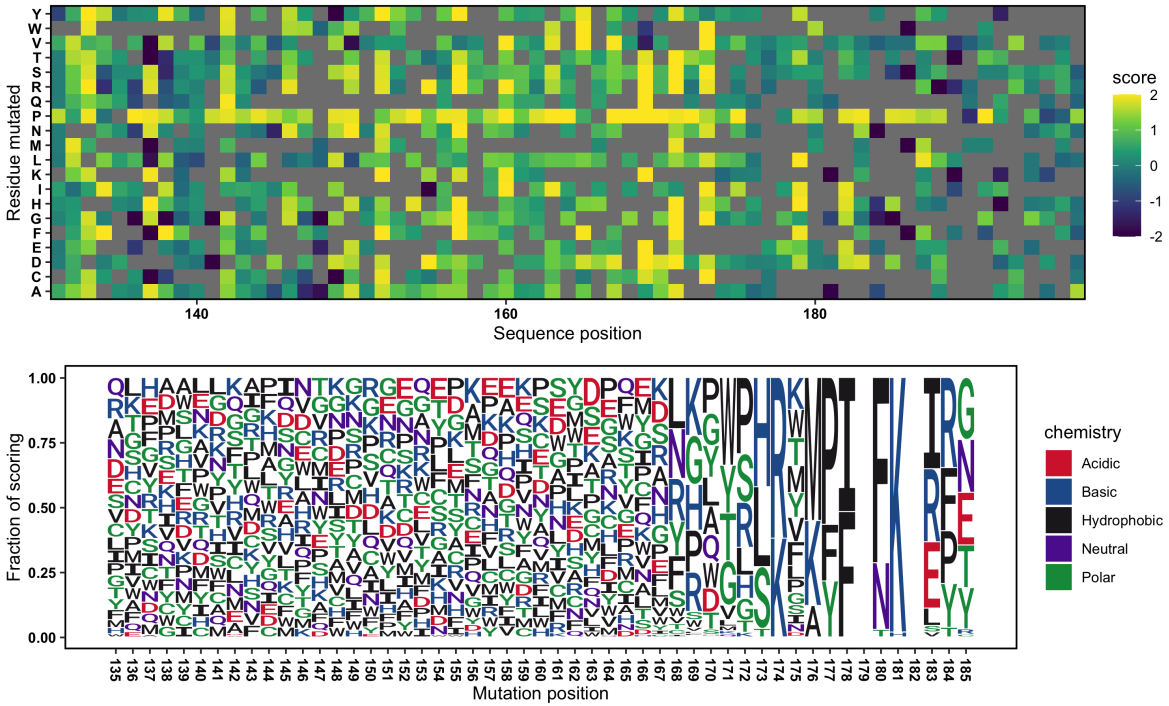
	Protein	Target	Biological activity	Species	Num of variants	Score
Data set 1	BRCA1	BARD1 RING domain	Ubiquitin E3 activity	*H. sapiens*	5610	Y2H assays
Data set 2	ERK2	Small molecule (SCH772984)	Resistance to drugs	*H. sapiens*	6810	Drug sensitivity assays. Calculation of cell availability
Data set 3	LDLRAP1	CBFC1	Protein translation	*H. sapiens*	6385	Y2H assays
Data set 4	Pab1	eI4FG1	Translation initiation	*S. cereviseae*	1340	Y2H assays

Example data set 4 - data overview

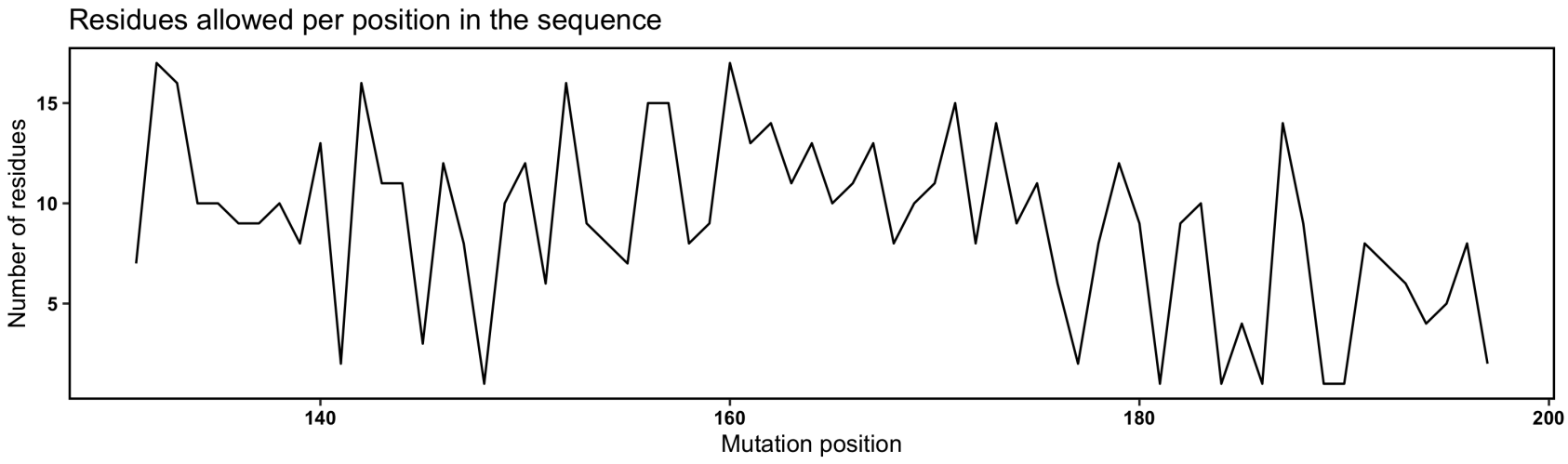


Score density plof scores per residue mutated

Example data set 4 - heatmap



Example data set 4 - conserved regions



Machine learning toolbox

- Ideas for supported machine learning framework:
  - Gaussian Process Regression.
  - Artificial Neutral Network.
  - ElasticNet Regression.

Results

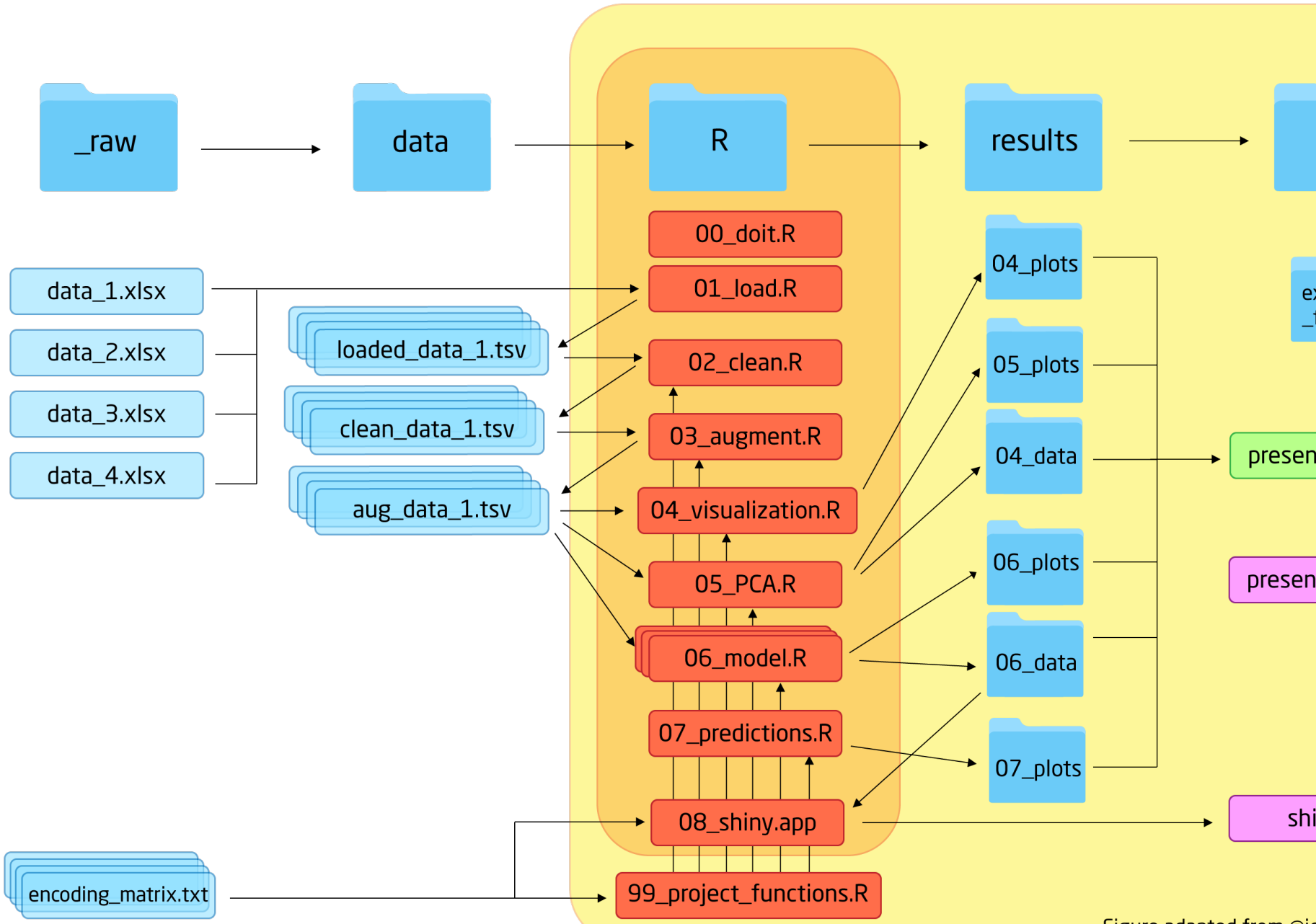
Discussion

References

- **Data set 1**: L. M. Starita, D. L. Young, et al. *Massively*

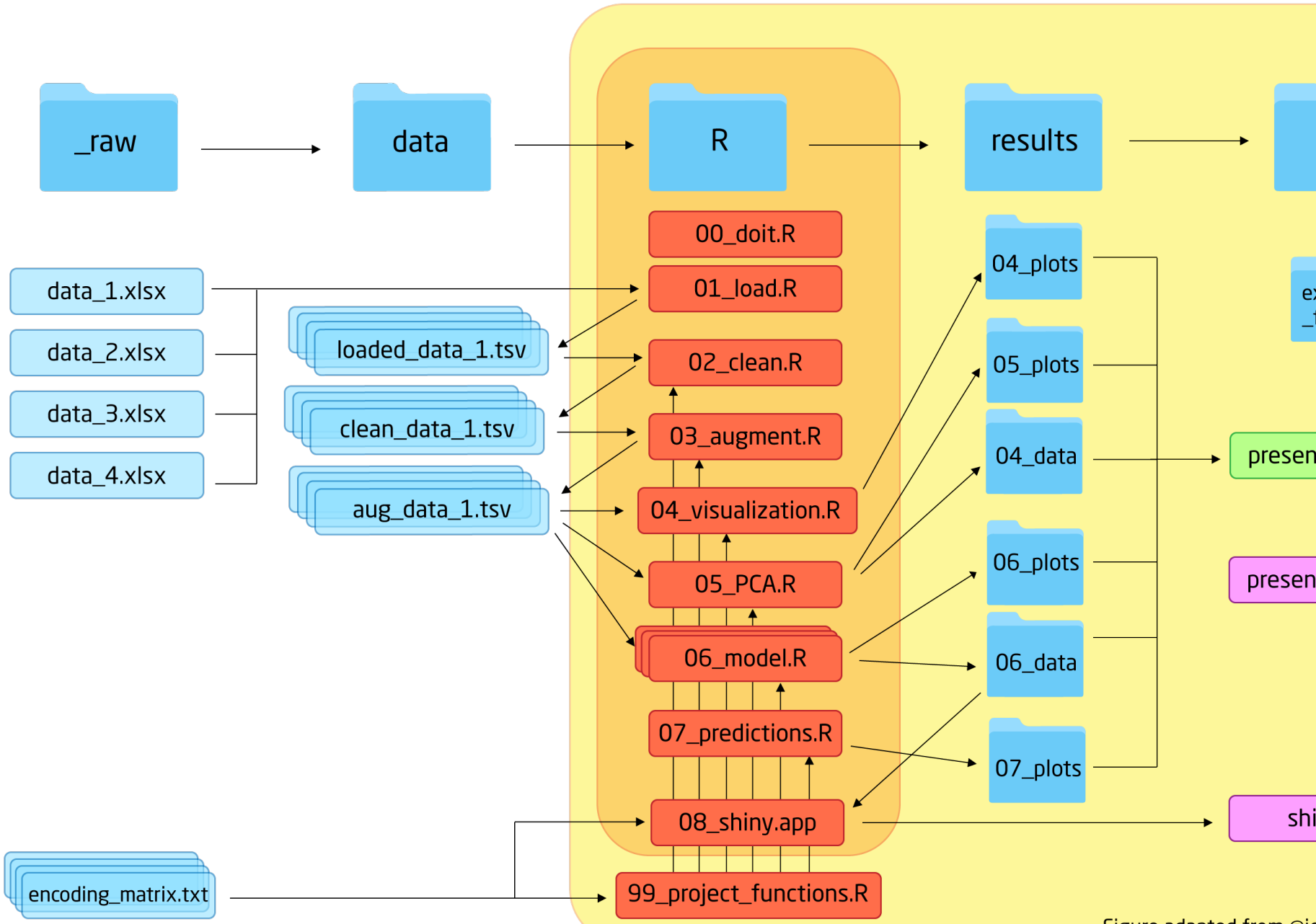
Appendix

R script overview 1



Appendix

R script overview 2



Appendix

R script overview 3