

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
- ▶ Conclusion

Introduction

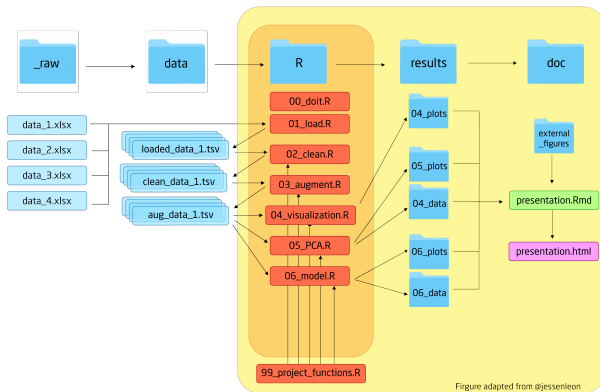
Prediction of protein-protein interactions (PPI) are a challenging task.

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

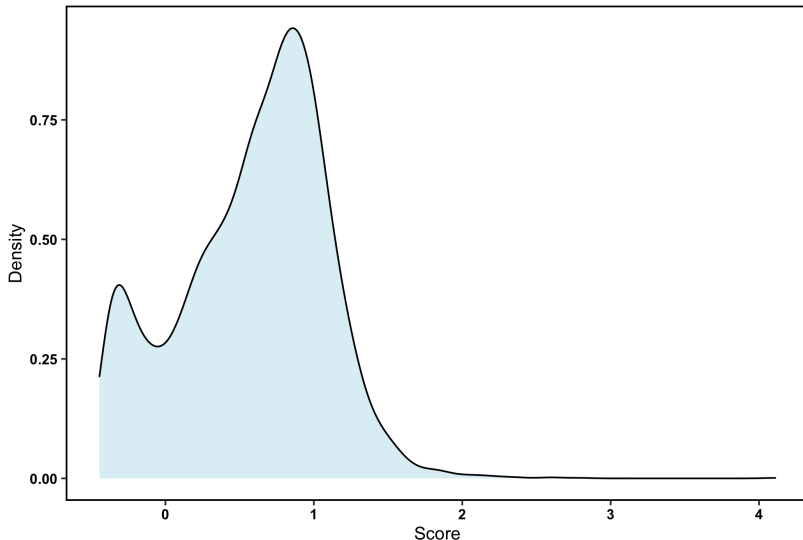
```
library("tidyverse")  
library("dplyr")  
library("devtools")  
library("Peptides")  
library("ggseqlogo")  
library("UniprotR")  
library("neuralnet")  
library("keras")
```

Methods - the data sets

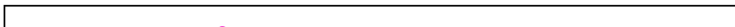
	Protein	Target	Biological activity	Species	Num of variants	Score
Data set 1	BRCA1	UBA1 RING domain	Ubiquitin E3 activity	<i>H. sapiens</i>	5610	Y2H assays
Data set 2	ERK2	Small molecule (SCH772984)	Resistance to drugs	<i>H. sapiens</i>	6810	Drug sensitivity assays. Calculation of cell availability
Data set 3	LDLR	ACE1	Protein translation	<i>H. sapiens</i>	6385	Y2H assays
Data set 4	Pab1	elF4G1	Translation initiation	<i>S. cerevisiae</i>	1340	Y2H assays

Example data set 1 - data overview

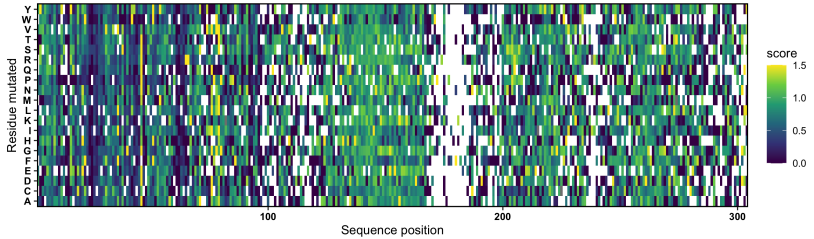
Score density plot for Data Set 1



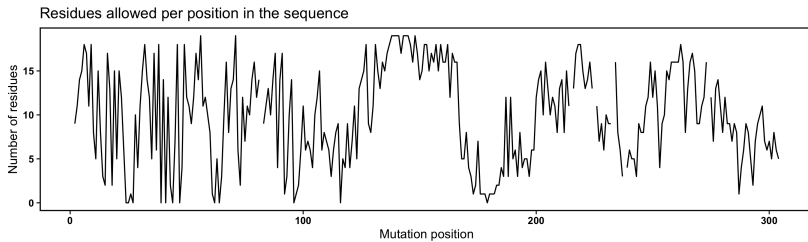
Score density plof scores per residue mutated



Example dataset 1 - heatmap



Example dataset 1 - conserved regions



Machine learning toolbox

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

Results

Conclusion

Referencex



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

Appendix

R script overview 1

```
data_set_1 <- read_tsv(file = './data/_raw/loaded_data_1.tsv')
```

```
# Data_set 1
```

```
# select variant_ID and activity to predict
```

```
data_set_1_clean <- data_set_1 %>%
```

```
  select(Variant_ID, E3_score) %>%
```

```
  rename(
```

```
    variant = Variant_ID,
```

```
    score = E3_score
```

```
  ) %>%
```

```
  filter(!str_detect(variant, "\\\\"))
```

► 02_clean.R

- Load: Load data from 01_load.R

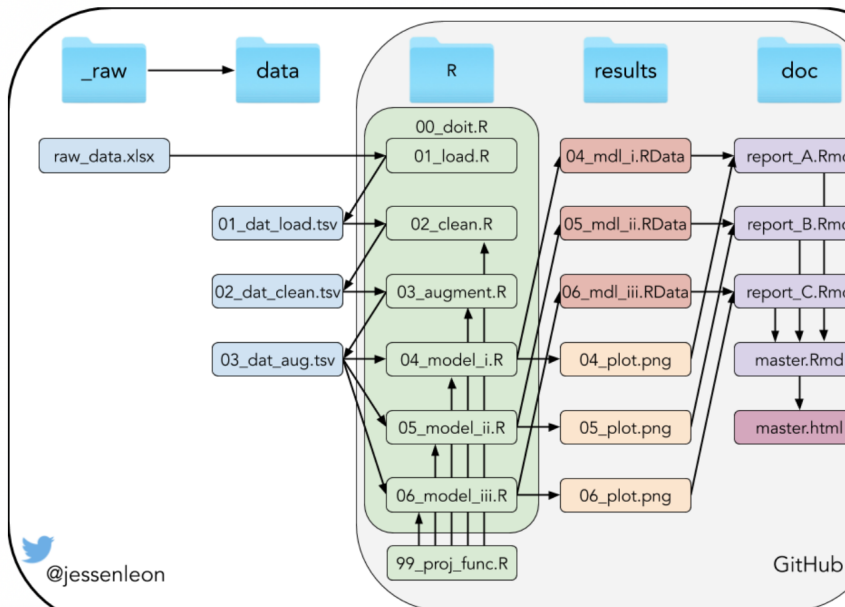
- Wrangle data: Remove NaN, fixes

- Save cleansed data in .tsv format

► 03_augment.R

Appendix

R script overview 2



Appendix

R script overview 3

