Machine learning EDA Report

Dennis Haandrikman

05/10/2022

Contents

Introduction

xx

Methods

xx

Results

Looking at the imported data, we can see that there are 8 different protein/peptide properties, these were explored further and shown below.

```
pander(head(bcel_sars_report[,6:13]), caption = "Header of the peptide/protein properties")
```

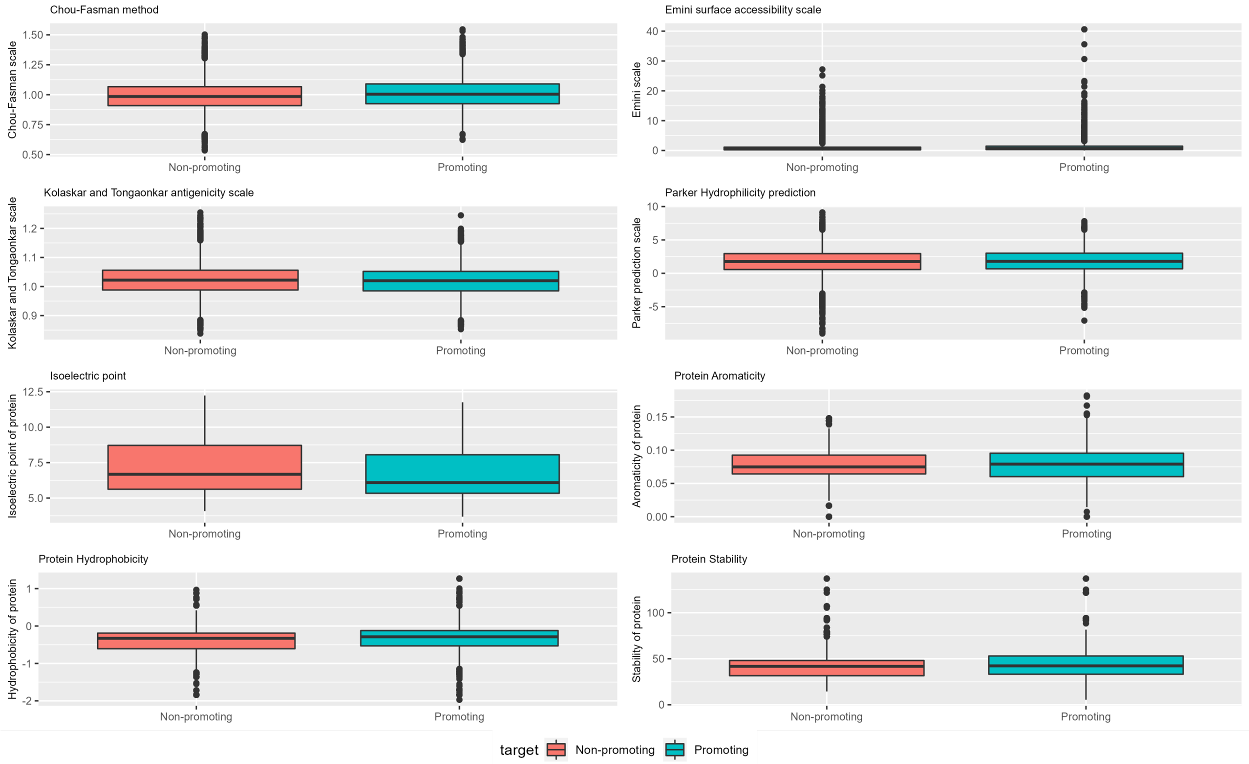
Table 1: Header of the peptide/protein properties (continued below)

chou_fasman	emini	kolaskar_tongaonkar	parker	isoelectric_point	aromaticity	hydrophobicity	stability
1.016	0.703	1.018	2.22	5.81	0.1033	-0.1438	40.27
0.77	0.179	1.199	-3.86	6.211	0.06548	-0.03691	25
0.852	3.427	0.96	4.28	8.224	0.09179	0.8792	27.86
1.41	2.548	0.936	6.32	4.238	0.04478	-0.5214	30.77
1.214	1.908	0.937	4.64	6.867	0.1038	-0.5788	21.68
0.928	0.547	1.09	0.9	6.867	0.1038	-0.5788	21.68

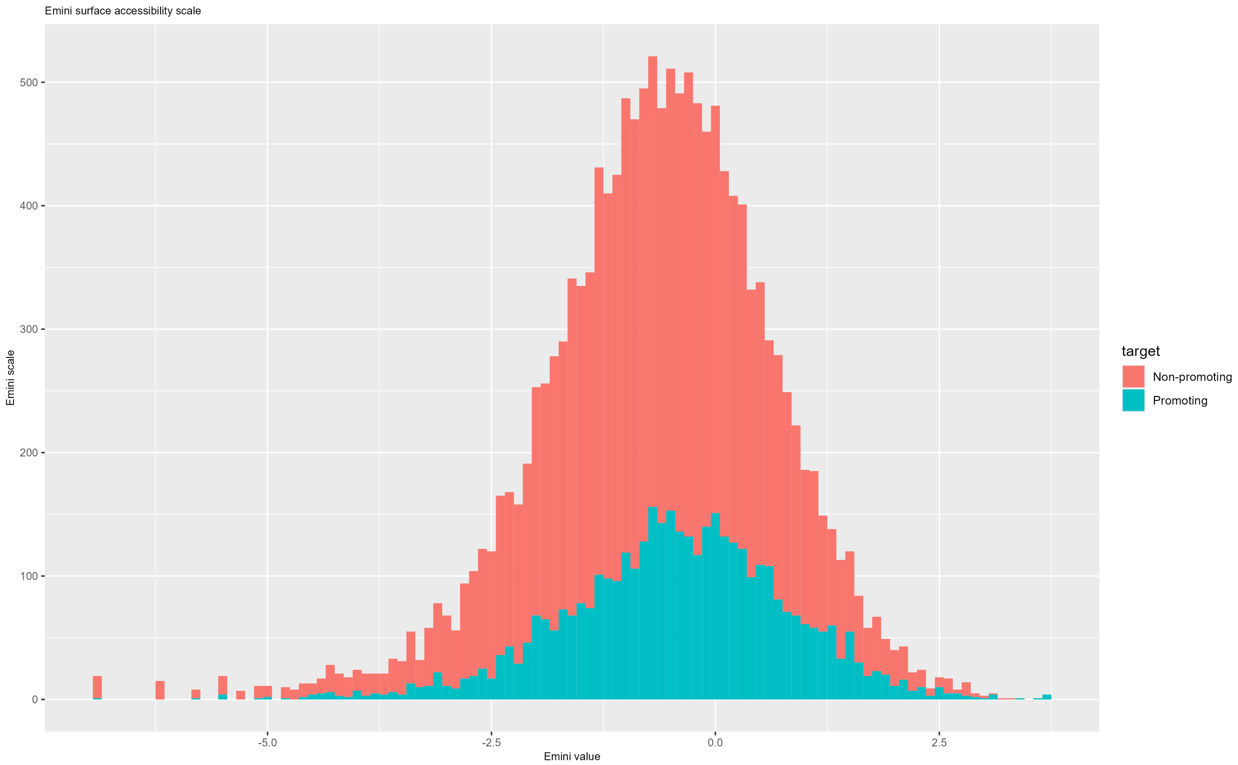
As we can see in the table, the first 4 columns; Chou_fasman, emini, kolaskar_tongaonkar & parker are all peptide properties. Where as the last 4 columns; isoelectric_point, aromaticity, hydrophobicity & stability are all protein properties.

The protein properties are from the full protein the peptide sequence is derived from.

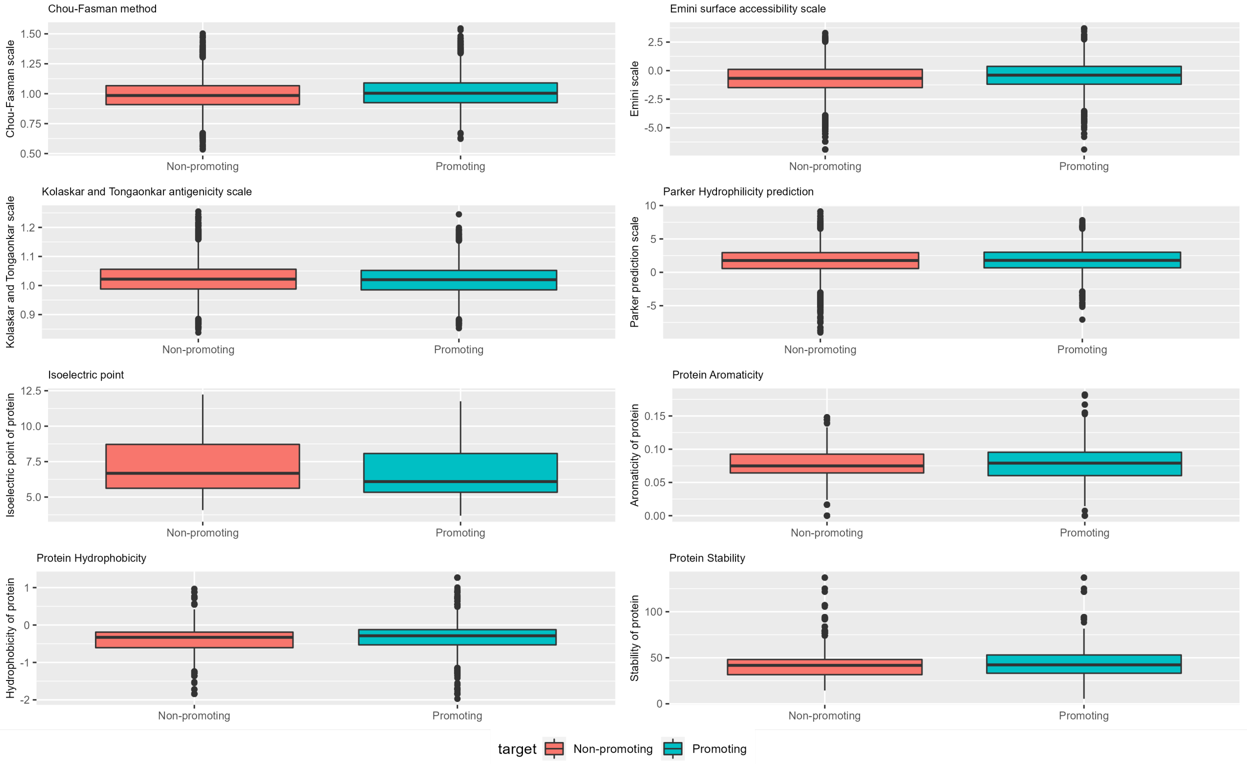
As such a boxplot has been made to showcase the properties of the peptide/protein properties, these have been grouped by the fact that they promote or don't promote(non-promoting) antigen binding sites.


Chou-Fasman method
Chou-Fasman scale
1.50
1.25
1.00
0.75
0.50
Non-promoting
Promoting
Emini surface accessibility scale
Emini scale
40
30
20
10
0
Non-promoting
Promoting
Kolaskar and Tongaonkar antigenicity scale
Kolaskar and Tongaonkar scale
1.2
1.1
1.0
0.9
Non-promoting
Promoting
Parker Hydrophilicity prediction
Parker prediction scale
10
5
0
-5
Non-promoting
Promoting
Isoelectric point
Isoelectric point of protein
12.5
10.0
7.5
5.0
Non-promoting
Promoting
Protein Aromaticity
Aromaticity of protein
0.15
0.10
0.05
0.00
Non-promoting
Promoting
Protein Hydrophobicity
Hydrophobicity of protein
1
0
-1
-2
Non-promoting
Promoting
Protein Stability
Stability of protein
100
50
0
Non-promoting
Promoting
target
Non-promoting
Promoting

As we can see in the figure, the boxplot for the emini data is unreadable and the data for that plot has been explored further. In the end it was necessary to perform a log transformation on the data as simply removing the outliers in the plot caused a loss of usable data. The plot of the histogram for the log transformed data is below as follows.

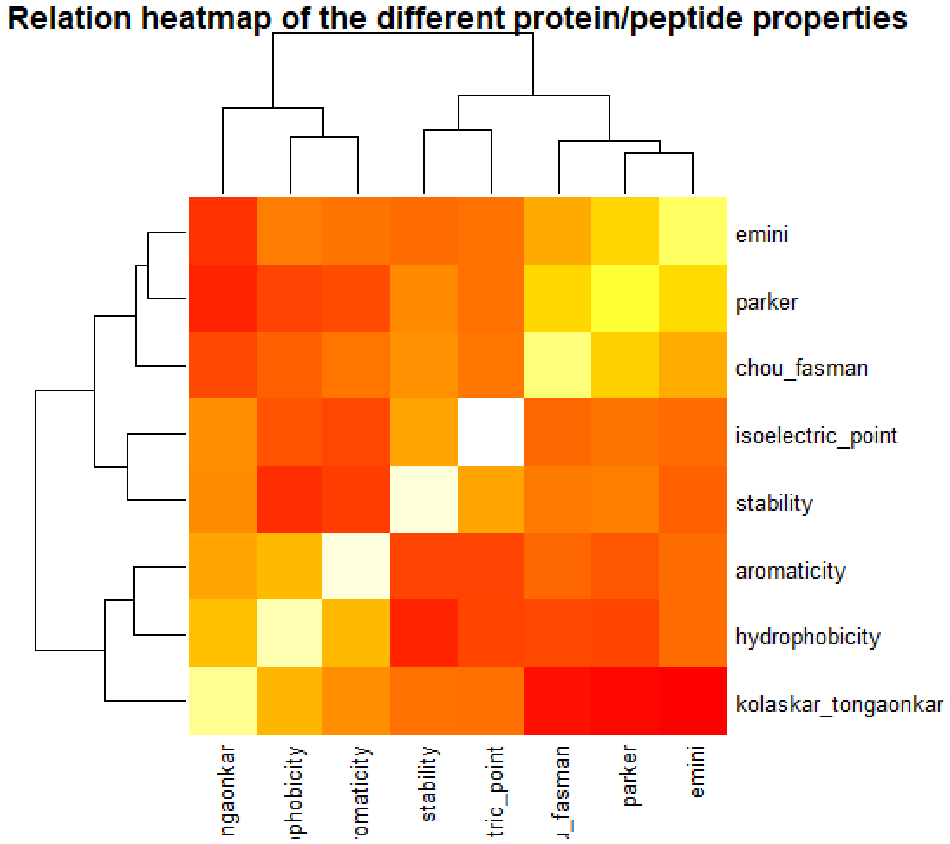


As we can see in the histogram data of the log transformed Emini data, we can see that it's now more normally distributed than in the boxplot. As such we can confirm that the log transformation of the Emini data is a success, to further prove said point all boxplots had been plotted again, but with the Emini data being the log transformed data, see the image below.



As we can see in the new boxplots is that the Emini data is now normally distributed, albeit with outliers lower and higher still. This further proves the point that the Emini data needs to be log-transformed to be usable further on in the machine learning process.

To continue to prepare for the machine learning process it would be wise to see what the relation of the peptide/protein properties are in regard to each other. For that a correlation matrix was calculated and plotted into a heatmap which follows bellow.



As we can see from the plotted correlation matrix heatmap, some aspects of the data are indeed related to each other, while some other aspects, surprisingly aren't. This might be due to the different scales and ranges that are being used for each of the properties.

Now that we've looked at the correlations between the properties, it's time to perform a Principal Component Analysis to visualize the variation in the dataset, we'll further see the correlation between the properties as well if done so.

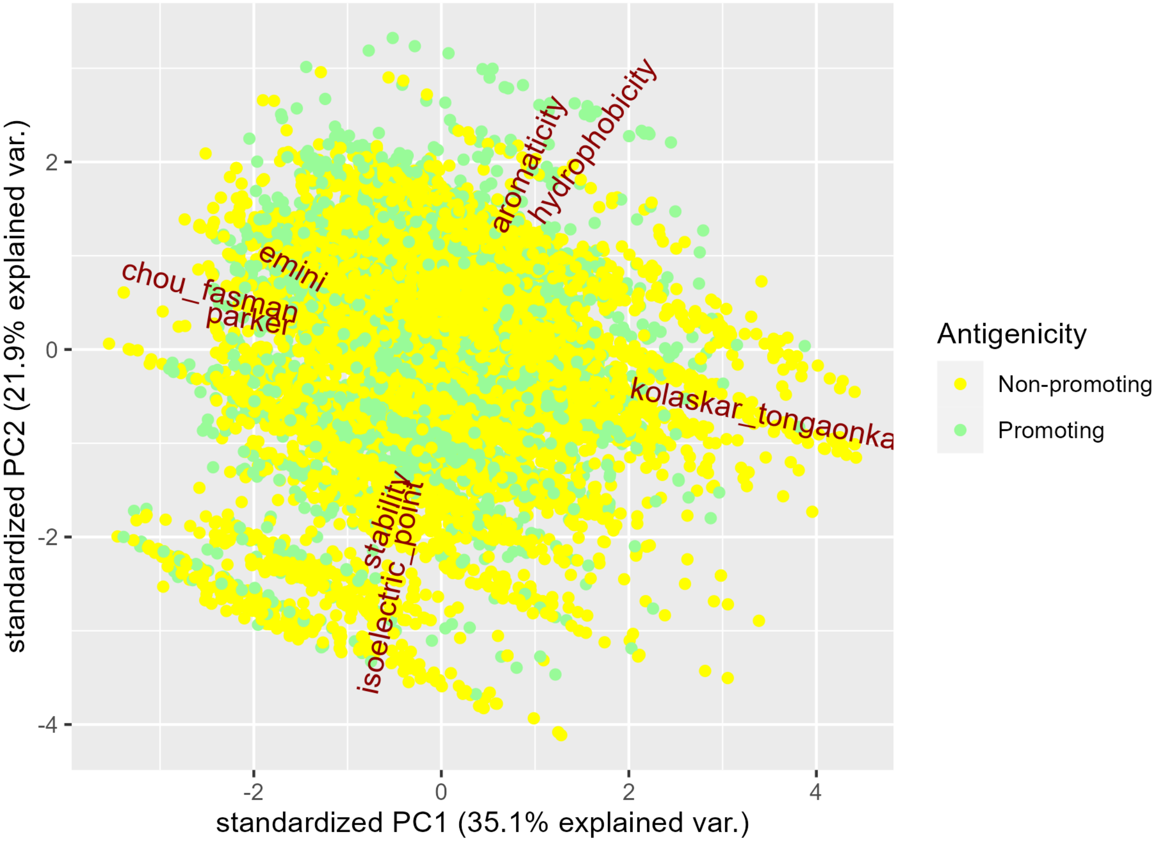


Figure 1: PCA plot of the protein/peptide properties

As we can see from the PCA is that some variables are closely interlinked, forming 4 different groups. These being: 1. Chour_fasman, Parker & Emini 2. Aromaticity, hydrophobicity 3. stability, isoelectric_point 4. Kolaskar_tongaonkar

And further we can see that there aren't really any clear differences between promoting and non-promoting peptide sequences, which might prove trouble-some for the machine learning algorythm, the data might need to be cleaned/transformed/standardized more.

Discussion and Conclussion

From the initial EDA exploration, we can see that the data consists of different scales for the peptide sequences. Which is worrying if we wish to compare those.

That was further proven by the correlation heatmap and the PCA plot, some properties are related to each other despite not only belonging to peptides or proteins. As we can see with the groups that have formed in the PCA plot, explained in the results above.

In the end, a quick conclusion is that the data needs to either be cleaned/transformed or standardized.