RBD Transmissibility Research

Presentation By Jake Roggenbuck
5/5/2022

RBD Finder (Python Script)

Abstract: A tool to align and find a subsequence of amino acids or nucleotides, when given a vaguely similar reference sequence.

Implementation: Used to find RBD from the spike of sars-like viruses.

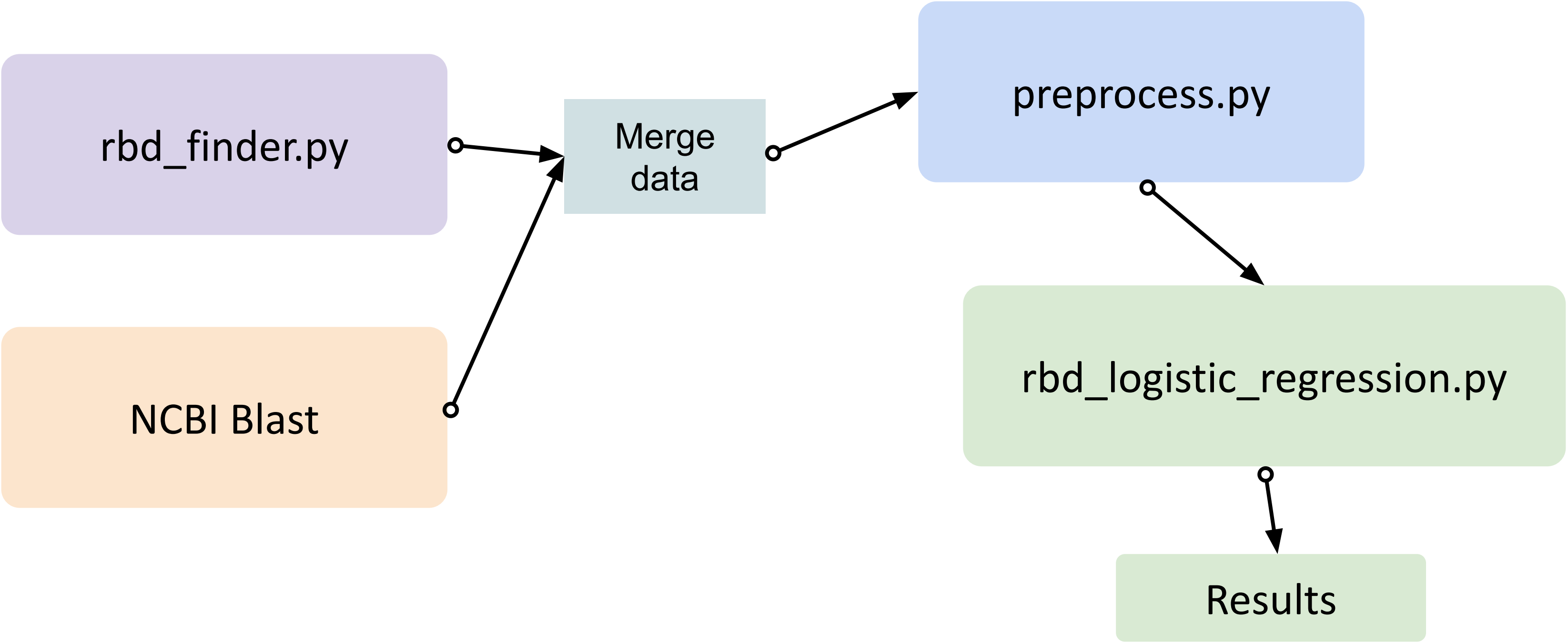
Why the RBD? The thinking behind using the RBD is that it could lead to less possible overfitting of the model (reducing possible noise from the spike).

Remove duplicates: df = df.drop_duplicates(subset=['Accession'])

How it works

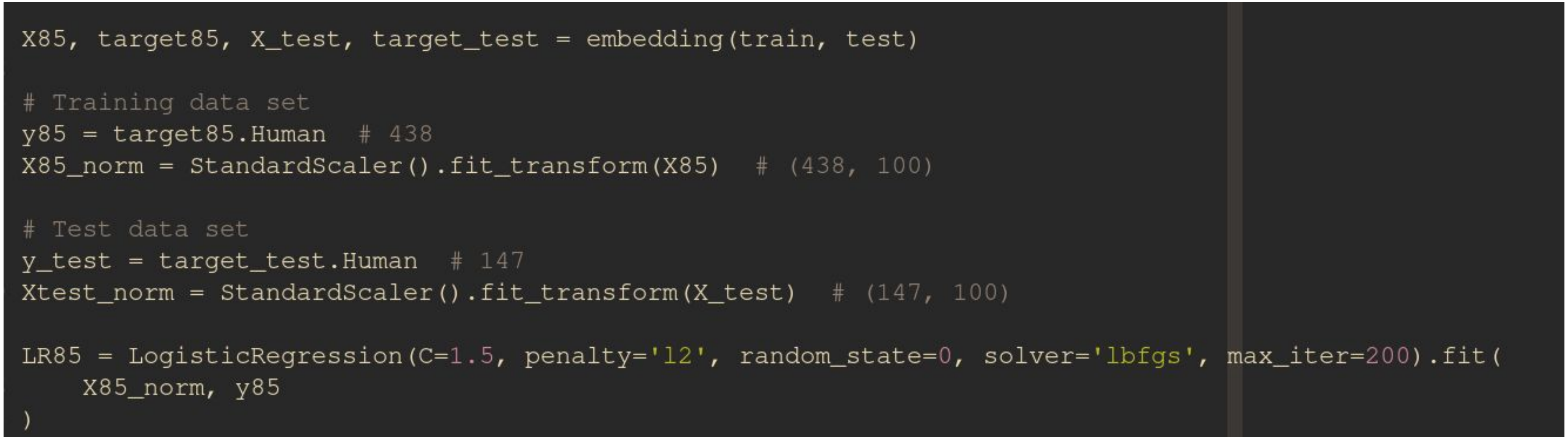
```
Original = “MKFLLLLSLFFPLSCAQDFSCNGHQTDTMALLRLNL”
Reference = “SLFFPLSCAADFCCNGHQTD”


Output = “MKFLLLLSLFFPLSCAQDFSCNGHQTDTMALLRLNL”
          0 ---- 7 ---------------- 26 ----- 35


Full = (MKF...NL, SLFFPLSCAQDFSCNGHQTD, 7, 26)
        Original, Extracted RBD,   Start, End
```

Process using rbd_finder.py



Logistic Regression

- How do we feed data to the regression (embedding)
  - Word2vec
  - Trigrams



```
X85, target85, X_test, target_test = embedding(train, test)

# Training data set
y85 = target85.Human  # 438
X85_norm = StandardScaler().fit_transform(X85)  # (438, 100)

# Test data set
y_test = target_test.Human  # 147
Xtest_norm = StandardScaler().fit_transform(X_test)  # (147, 100)

LR85 = LogisticRegression(C=1.5, penalty='l2', random_state=0, solver='lbfgs', max_iter=200).fit(
    X85_norm, y85
)
```

RBD Extraction Retrospective

The RBD finder script was not always able to identify the RBD for every sequence, so sometimes manual intervention needed to be done to find the sequence using NCBI's BLAST.

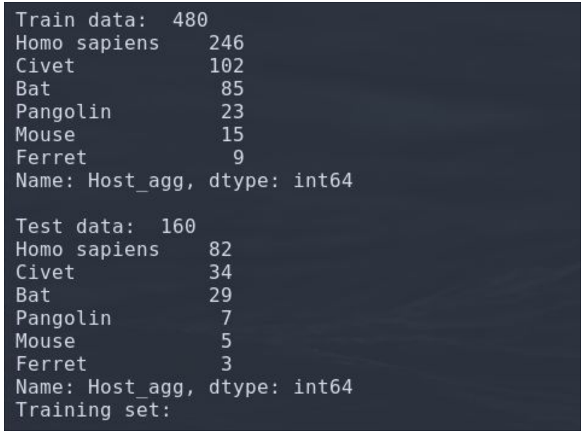
The RBD finder has been able to find up to 100% of the RBD in certain samples however the average it would find was around 95% for most samples. Requiring a user to manually find the last 5% using BLAST.

SARS1 Any (Full Spike) (Note that this includes MERS)

ROC_AUC Accuracy = **0.98**

Sample Size = **640**

98%



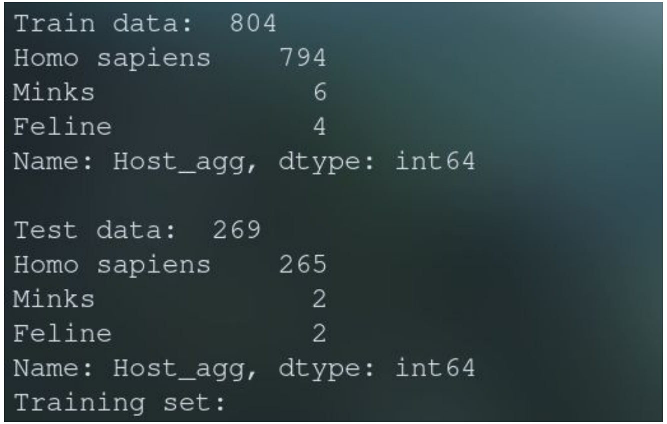
```
Train data:  480
Homo sapiens    246
Civet           102
Bat              85
Pangolin         23
Mouse            15
Ferret            9
Name: Host_agg, dtype: int64

Test data:  160
Homo sapiens    82
Civet           34
Bat             29
Pangolin         7
Mouse            5
Ferret           3
Name: Host_agg, dtype: int64
Training set:
```

Subset of just SARS-CoV-2 (Full Spike)

ROC_AUC Accuracy = **0.67**

Sample Size = **1073**

67%



```
Train data:  804
Homo sapiens    794
Minks             6
Feline            4
Name: Host_agg, dtype: int64

Test data:  269
Homo sapiens    265
Minks             2
Feline            2
Name: Host_agg, dtype: int64
Training set:
```

SARS1 Any (RBD)

96%

ROC_AUC Accuracy = **0.96**

Sample Size = **640**

Subset of just SARS-CoV-2 (RBD)

98%

ROC_AUC Accuracy = **0.98**

Sample Size = **1165**

Summary

	Full Spike	RBD
SARS1 Any	98%	96%
SARS-CoV-2	67%	98%

* In ROC Accuracy

False Negatives

Usually when there is missing data near the RBD

MF…PLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTXXXXXXXXXXXXXXXXXXXXXXXXXXTDAVDCALDPLSETKCTLKSFTVE

MF…VFLVEGFNCXFPLQSYGFQPTNXVGXXPXXXXXXXXXXXXXXXXXXXXXKSTNLVKNKCVNFNFNGLTGTGVLTESN

This is another reason to use BLAST because of its sequence searching.

Conclusion

- Logistic Regression with RBD
  - Sometimes improved ROC accuracy in the tests conducted
  - Was not shown to significantly reduce ROC accuracy
- RBD finder
  - Sometimes needs manual intervention for finding missing sequences
- Missing Data
  - False Negatives are from missing data in or near the RBD