VIREX

Gözde Gökyokuş

Supervised by Asst. Prof. İlknur Karadeniz

IŞIK UNIVERSITY
COMPUTER SCIENCE AND ENGINEERING

Introduction

As the world is facing a dangerous pandemic nowadays, we clearly realize the importance of medical services and the importance of benefitting computational methods in order to get to know different life forms and their interactions. Thanks to recent technological developments in bioinformatics, analyzing and interpreting the biological data is now faster and more accurate. Identifying the different characteristics and structures of the human proteins and pathogen proteins that infect us, become essential to be more proactive in humans fight with pathogens. Protein sequence analyzing makes it possible to predict various potential interactions between virus and human proteins by identifying the significant features and classifying them with respect to their similarities and differences.

In this project, Virex, the objective is to develop a model in order to predict possible infections between human and virus proteins through gathering protein sequences by web scraping, determining the main features of those proteins and classifying the various interactions by using machine learning algorithms.

The developed model consists of three subsystems: Data Gathering, Feature Engineering, and Prediction. While Data Gathering subsystem is responsible from extracting the protein sequence data from the web, Feature Engineering subsystem converts that data into required form and identifies the significant features in predicting the interactions between human and virus proteins which is accomplished by Prediction subsystem.

The model is tested with different data sizes and using different prediction algorithms. The results obtained with the proposed model, Virex, are really promising as 0.97 accuracy rate is obtained, outperforming the accuracy values provided in the literature.

Implementation Details

Virex is implemented in **Python language** on **Google Colab environment**.

Initial Pathogen-Host Interaction data was downloaded from **Phisto DB web page** then imported into the project with the help of **Pandas** library. Data visualizations are made with **Matplotlib** module.

Custom URLs are created with the path of unique virus and human proteins ids. With the help of **Urllib** library, the sequence data on **Uniprot.org** is extracted. In order to decrease the execution time in web scraping, **multiprocessing** library is used and 5 different processes work at the same time. Hardware –software mapping is visualized as it can be seen on Figure A. The resulting performance upgrade can be seen on Table 1 which was calculated with **time** library. Then human and virus sequences are **cleaned** from the unnecessary characters and saved into a CSV file. Some of the sequence data which are not available on the web is written on a txt file with **pickle** library to be excluded on all records later on. Track of the infection between the virus and human proteins are kept with the IDs on the initial data so the gathered sequences are joined according to those IDs. **Cartesian product** between the unique human and virus proteins are created with **Itertools** library just for creating the infection negative records. By comparing with the positive infection records negative infections are labeled as Infection False. Infection column is stated as the target column, Virus Sequence and Human Sequence columns are stated as the features.

Amino acid sequences are split into their amino acids and counts of the characters are listed as a matrix by using **Count Vectorizer**. Then by using **Tf-idf Transformer**, their term frequencies and inverse document frequencies are calculated and their product is written on the related columns. The frequencies of matching virus and human amino acids are added together and divided into 2 in order to have their vectoral frequency representation as a match. Infection column is binarized by using **Label Binarizer**.

Features and labels are split into training and test sets with **Sklearn** library (0.2 rate) For experimental purposes, Virex applies **different portions of virus-host match data** into several ML algorithms to predict the infection between the virus and host proteins.

Those algorithms are as follows:

-Random Forest Classifier

-Naïve Bayes Classifier

-Support Vector Classifier

The classification algorithms, confusion matrices, classification reports are implemented with Sklearn library.

Execution Time	Technique
0.20561695098876953 seconds	Saving and reading from files
788.8343715667725 seconds	Web scraping with urllib (Multiprocessing)
4148 seconds	Web scraping with urllib

Table 1. Execution times according to the data collecting method

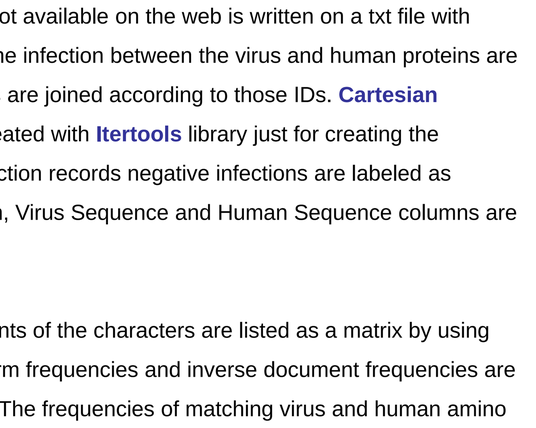

Figure A. HW-SW Relations

Results

```
         HumanProtein VirusProtein  \
0              P15311       Q9WIP8
1              P28799       Q6JN72
2              P62937       P88761
3              P62937       P88585
4              P23284       Q77YG1
...               ...          ...
11238988       K7PPA8       Q80089
11238989       K7PPA8       P00036
11238990       K7PPA8       Q88019
11238991       K7PPA8       O40955
11238992       K7PPA8       Q3ZK63

                                                VirusSequence  \
0         MGGKWSKSSIIGSPTVRERMKRAEPAADGVGAVSRDLDKYGAITSS...
1         MEPVDPSLEPWKHPGSQPKTACTNCYCKKCCLHCQVCFHKKGLGIF...
2         PIVQNLQRQIVHQALSPRTLNAWVKVVEEKAFSPEVIPMFSALSEG...
3         ELERFAVNPGLLETSEGCRQILGQLQPSLQTGSEELRSLFITVAVL...
4         MGARASVLSGGELDRWEKIRLRPGGKKKYKLKHIVWASRELERFAV...
...                                                     ...
11238988  MVPQALLFVPLLVFPLCFGKFPIYTIPDKLGPWSPIDIHHLSCPNN...
11238989  MDRVLSRADKERLLELLKLPRQLWGDFGRMQQAYKQQSLLLHPDKG...
11238990  MGGAISMRRSRPSGDLRQRLLRARGETYGRLLGEVEDGYSQSPGGL...
11238991  MERLLDEVLAPGGPYNLTVGSWVRDHVRSIVEGAWEVRDVVSAAQK...
11238992  MLKMESTQQMASSIINTSFEAAVVAATSTLELMGIQYDYNEVVTRV...

                                                HumanSequence Infection
0         MPKPINVRVTTMDAELEFAIQPNTTGKQLFDQVVKTIGLREVWYFG...      True
1         MWTLVSWVALTAGLVAGTRCPDGQFCPVACCLDPGGASYSCCRPLL...      True
2         MVNPTVFFDIAVDGEPLGRVSFELFADKVPKTAENFRALSTGEKGF...      True
3         MVNPTVFFDIAVDGEPLGRVSFELFADKVPKTAENFRALSTGEKGF...      True
4         MLRLSERMMKVLLAAAALIAGSVFFLLLPGPSAADEKKKGPKVTVKV...      True
...                                                     ...       ...
11238988  MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLS...     False
11238989  MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLS...     False
11238990  MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLS...     False
11238991  MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLS...     False
11238992  MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLS...     False

[11238993 rows x 5 columns]
```

Figure 1. 11 Million records of virus-host sequence matches

```
                a         c         d         e         f         g         h
0        0.275332  0.048503  0.229607  0.429746  0.127160  0.224107  0.121395
1        0.200314  0.414100  0.202211  0.170795  0.077886  0.254461  0.141532
2        0.309299  0.047461  0.136971  0.307771  0.213429  0.377830  0.123963
3        0.237064  0.082497  0.130874  0.322872  0.209259  0.365640  0.064423
4        0.271197  0.051762  0.186108  0.267286  0.185050  0.370100  0.068266
...           ...       ...       ...       ...       ...       ...       ...
499995   0.289536  0.068415  0.139966  0.281287  0.139598  0.377536  0.104060
499996   0.235568  0.130765  0.187395  0.203510  0.091736  0.372346  0.050204
499997   0.307043  0.079457  0.203418  0.287085  0.157791  0.397131  0.059176
499998   0.284705  0.101579  0.164418  0.319286  0.180887  0.402547  0.099710
499999   0.224623  0.089759  0.223220  0.316030  0.216241  0.389128  0.078611

                i         k         l         m         n         p         q
0        0.168187  0.319880  0.366639  0.118421  0.125670  0.209022  0.209223
1        0.043279  0.324645  0.177707  0.061014  0.064684  0.357196  0.235056
2        0.210684  0.218755  0.229999  0.132727  0.199319  0.239631  0.184982
3        0.217235  0.297382  0.247043  0.091158  0.189038  0.137350  0.269165
4        0.202286  0.373062  0.275522  0.108491  0.151190  0.205374  0.156774
...           ...       ...       ...       ...       ...       ...       ...
499995   0.106573  0.301023  0.284526  0.061737  0.133725  0.302345  0.164608
499996   0.084837  0.335617  0.390599  0.081941  0.097470  0.213040  0.244480
499997   0.111761  0.217896  0.318436  0.090220  0.134985  0.275261  0.175791
499998   0.157836  0.224997  0.269142  0.072733  0.146191  0.282163  0.183165
499999   0.140604  0.313632  0.395111  0.081296  0.144821  0.173445  0.148230

                r         s         t         v         w    x         y    u
0        0.241958  0.197145  0.147688  0.189830  0.089219  0.0  0.133386  0.0
1        0.229113  0.280551  0.196320  0.159718  0.053920  0.0  0.055639  0.0
2        0.163014  0.185542  0.220939  0.271621  0.052927  0.0  0.023423  0.0
3        0.136126  0.241863  0.244229  0.236800  0.012834  0.0  0.057393  0.0
4        0.204282  0.195375  0.230316  0.282183  0.058514  0.0  0.076483  0.0
...           ...       ...       ...       ...       ...  ...       ...  ...
499995   0.247287  0.270572  0.212088  0.205534  0.034561  0.0  0.105679  0.0
499996   0.199087  0.224902  0.176906  0.213378  0.076560  0.0  0.101891  0.0
499997   0.280130  0.193710  0.192485  0.219102  0.055086  0.0  0.110304  0.0
499998   0.224303  0.207255  0.241061  0.219702  0.030844  0.0  0.121948  0.0
499999   0.178723  0.176062  0.165571  0.201813  0.075524  0.0  0.101235  0.0

[500000 rows x 22 columns]
```

Figure 2. Tf-idf value averages of Human and Virus Sequence Vectors

```
[[11105   870]
 [ 2019  6006]]
              precision    recall  f1-score   support

           0       0.85      0.93      0.88     11975
           1       0.87      0.75      0.81      8025

    accuracy                           0.86     20000
   macro avg       0.86      0.84      0.85     20000
weighted avg       0.86      0.86      0.85     20000

0.85555
```

Figure 3. Prediction of 100.000 rows of data on Support Vector Classifier

```
[[11975     0]
 [ 8022     3]]
              precision    recall  f1-score   support

           0       0.60      1.00      0.75     11975
           1       1.00      0.00      0.00      8025

    accuracy                           0.60     20000
   macro avg       0.80      0.50      0.37     20000
weighted avg       0.76      0.60      0.45     20000

0.5989
```

Figure 4. Prediction of 100.000 rows of data on Naïve Bayes Classifier

```
[[1798  282]
 [  58 7862]]
              precision    recall  f1-score   support

           0       0.97      0.86      0.91      2080
           1       0.97      0.99      0.98      7920

    accuracy                           0.97     10000
   macro avg       0.97      0.93      0.95     10000
weighted avg       0.97      0.97      0.97     10000

0.966
```

Figure 5. Prediction of 50.000 rows of data on RF Classifier

```
[[11677   298]
 [  911  7114]]
              precision    recall  f1-score   support

           0       0.93      0.98      0.95     11975
           1       0.96      0.89      0.92      8025

    accuracy                           0.94     20000
   macro avg       0.94      0.93      0.94     20000
weighted avg       0.94      0.94      0.94     20000

0.93955
```

Figure 6. Prediction result of 100.000 rows of data on RF Classifier

```
[[191929    214]
 [  5277   2580]]
              precision    recall  f1-score   support

           0       0.97      1.00      0.99    192143
           1       0.92      0.33      0.48      7857

    accuracy                           0.97    200000
   macro avg       0.95      0.66      0.74    200000
weighted avg       0.97      0.97      0.97    200000

0.972545
```

Figure 7. Prediction of 100.000 rows of data on RF Classifier

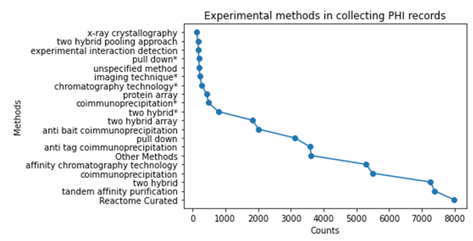

Figure 8. Experimental methods used for interaction observation in PHISTO data

Conclusion

This project aims at predicting human and virus protein interactions proposing machine learning based models. The developed system composed of three subsystems being Data Gathering, Feature Engineering, and Prediction is implemented with various data sets and the results show that Virex is capable of successfully finding and extracting data from the internet, processing the data through feature engineering and predicting the possible infections.

-RF Classifier: Best results are performed with RF algorithm among all the other performed algorithms. If we evaluate different data portions, the accuracy of the classification was the highest with 1M row of data (0.97). 50K rows of data resulted as the next best accuracy (0.96). Then as the third experiment 100K rows of data resulted with an accuracy of 0.939.

For the 1M rows: 191929 of the infection negative values are guessed correctly, only 214 of them are wrong labeled as infection positive values. 2580 of the infection positive values are labeled correctly, 5277 of the positive infection values are mislabeled as negative infection.

For the 100K rows: 7114 of the infection positive values and 11677 of the infection negative values are classified correctly. 298 of the infection negative values are misclassified as infection positive, 911 of the infection positive values are misclassified as infection negative.

For the 50K rows: 7862 of the infection positive values, 1798 of the infection negative values are correctly labeled. 282 of the infection negative values are misclassified as infection positive, 58 of the infection positive values are misclassified as infection negative.

-SVM Classifier: Even though SVM was slightly underperforming than the RF classifier, despite that it gave a high accuracy which was 0.855. It guessed 11105 infection negative values and 6006 infection positive values correctly. Only it misunderstood the 870 of the infection negatives and 2019 of the infection positives.

-Naïve Bayes Classifier: NB algorithm was the worst in accuracy among all the other algorithms. From the confusion matrix and classification report it can be seen that the algorithm labeled 8022 of the infection positive values as infection negative values. That's why this accuracy is smaller than the others. But it guessed all of the infection negative values(11975), and 3 of the infection positive values correctly so the accuracy slightly increases. There are 0 mislabeled infection negative values.

Bibliography

1. Gozde Gokyokus, Thesis Report Virex, Işık University, Jan 2021
2. Sklearn. Count Vectorizer. tarih yok. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html>.
3. Label Binarizer. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelBinarizer.html>.
4. MultinomialNaiveBayes. <https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html>.
5. Train test split. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>
6. Tf-idf Transformer. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html>.
7. Google. Colaboratory. <https://research.google.com/colaboratory/faq.html>.
8. Pandas. tarih yok. <https://pandas.pydata.org/>.
9. Python Internet Access using Urllib.Request and urlopen. <https://www.guru99.com/accessing-internet-data-with-python.html>.
10. Python. Itertools. <https://docs.python.org/3/library/itertools.html>.
11. Multiprocessing. <https://docs.python.org/3/library/multiprocessing.html>.
12. Pickle. <https://docs.python.org/3/library/pickle.html>.
13. Time. <https://docs.python.org/3/library/time.html>.
14. Yiu, Tony. Random Forest. 12 06 2019. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>.
15. Matplotlib. <https://matplotlib.org/>.
16. UniProt. 24 05 2020. <https://en.wikipedia.org/w/index.php?title=UniProt&oldid=958490101 >.
17. Support vector machine. 20 10 2020. <https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=984421646 >.
18. Train test split. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
19. İlknur Karadeniz, Saliha Durmuş Tekir, Tunahan Çakır, Emre Ardıç, Ali Semih Sayılırbaş, Gökhan Konuk, Mithat Konuk, Hasret Sarıyer, Azat Uğurlu, Arzucan Özgür, Fatih Erdoğan Sevilgen, Kutlu Ö. Ülgen. «PHISTO: pathogen–host interaction search tool.» Bioinformatics 29 (2013): 1357-1358. https://doi.org/10.1093/bioinformatics/btt137 .
20. UniProt. 24 05 2020. <https://en.wikipedia.org/w/index.php?title=UniProt&oldid=958490101 >.
21. Kohli, Shivam. Understanding a Classification Report For Your Machine Learning Model. 18 11 2019. <https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>.