Host-Virus Protein-Protein Interaction Prediction and Analysis for novel Coronavirus Protein Strains

Anonymous CVPR submission

Paper ID

Abstract

Authors - Chandan Gupta
Sanskar Sachdeva
Sarthak Pal
(We were not able to ad authors at the right place due to some bug)
The current global pandemic caused by Covid-19 makes it imperative for biochemists and molecular biologists to study virus' disease causing interactions interactions with the host organism. Host-Virus protein-protein interactions, arising due to electrostatic forces among protein molecules have been extensively investigated for viral infections and have provided important biological insights as to how virus takes control of of the host cell. Currently these are investigated only through experimental techniques which are very time consuming and expensive. In this work, we present a machine learning algorithm which can predict whether an interaction exists between Sars-Cov2 and human protein strains. This model can be used to quickly test for interactions when new protein strains for the virus are discovered, and hence can be efficiently used by the scientists working in this area, as well as serve as a generalized tool for other viral infections.

1. Introduction

Proteins are the fundamental molecules responsible for all the biological processes required by an organism to sustain. Till the early 1980s, geneticists and structural biologists mainly studied individual proteins in isolation, ignoring that they may be mediating each other's functions. Studies started showing up suggesting that proteins rarely work alone, and instead worked collectively together, which can be realized by studying a set of highly mutually dependent proteins. This gave rise to new areas in biology, such as systems biology and network biology[1], former studying biological molecules as a system, while latter studying them as a network where each node is a bio molecule and the connections between them are based on certain factors like function similarity and dependence.

These developments made it possible to study how different virus' are able to take control of their host's cellular mechanism after entering. Control theory[2] has revealed that if we approximated this system as a network, the connections between virus' and host proteins are very few, suggesting that they target only certain nodes, called the hub or driver nodes. Hence, such systems studies have been very instrumental and has been promoted recently by many medical scientists. However, to construct such networks, exact protein-protein interactions (PPI) [3] have to be understood, and such data is recorded through experimental techniques like mass spectrometry. Such experiments are highly time confusing and expensive, which limits their practicality and PPI research. Therefore, it makes it extremely important to make reliable methods to identify PPI among different viruses and host proteins, which are fast, efficient and are cost effective, without compromising biological significance at the same time.

Machine learning algorithms are increasingly being used in biological applications, where deterministic and domain specific algorithms fail to work. Their capability to decipher deep biological insights from the data collected experimentally has been heaviliy exploited to understand the mechanisms behind certain biological processes, including PPIs. Therefore, in this paper, we propose to identify PPIs between Sars-Cov2 and homo sapiens with the help of a machine learning model, which can be efficiently be used as a software package or web server by scientists in the area to quickly check for PPIs when new protein strains are identified for the novel Sars-Cov2, instead of performing new experiments everytime.

2. Related Work

Machine Learning driven software have been proposed by many works in the PPI literature - random forests, ensemble learning, SVMs. [4,5,6] uses a family of viruses instead of a single virus, and uses doc2vec method for ex-

tracting biological significant features for building classifiers. [7] experimented with SVMs and Naive Bayes, and suggesting that Naive Bayes may not be a good model for this problem, validating that the biological interactions are not independent among two-tuples of virus and host proteins. [8] trained a Doc2Vec model for protein feature extraction from raw sequence, and has been popularly used in the community. Although the data provided by them is based on specific parameters - 100 length vector and 3 window slides, it can be quickly used to test a model. [9] used developed a different feature extraction method, known as a repeat pattern of amino acid composition, which considers maximum lengths of occurrences of specific motif in the sequence. This resulted in the best classification performance across the literature in this domain. Studies like [10] have reviewed the efficacy of machine learning and artificial intelligence (AI) techniques for detecting PPIs. Further, experimental research projects, described in [11] have built public repositories to help scientists develop classification solutions in this domain, thereby suggesting the need for the experimentalists to identify interactions for new sequences of a particular virus, after the rise of next generation sequencing (NGS).

3. Dataset and Preprocessing

Dataset was compiled from [12]. [12] contains sequences of host (Human) and three viruses of the coronavirus family - Sars-Cov1, Sars-Cov2, Mers-Cov, and whether a specific pair interacts or not. Since Sars-Cov2 is relatively new, not many protein sequences have been identified for it. [11] suggested that due to high sequence and homology similarity between the three viruses, they can be used for studying interactions. Therefore, inorder to have sufficient data, we also compiled interaction tuple sequences for the other two viruses as well. Total 736 tuple sequences were sampled from the resource. Each tuple had a number assigned to it - 0 if they interacted, 1 otherwise. Total number of positive samples (interacting) were 192, while negative samples (non interacting) were 544.

3.1. Feature Extraction

[4] suggested that Doc2Vec efficiently extracts features from raw biological sequences. Doc2Vec is used in Natural Language Processing (NLP) to extract feature vectors from document files to be fed into a learning model. It contains a shallow neural network, which is trained on a document to generate semantically significant feature vectors. [8] had stored a trained model, which generates for each 3-mer amino acid composition a feature vector of length 100, so that new models can be tested quickly by using this without repeatedly training a new Doc2Vec model on the same protein sequences. In this work, we used this model to extract 100 length vectors for both virus and human proteins, and then added them, as suggested by [8]. This makes our dataset suitable for a supervised machine learning problem, having a shape of (N,100), where N is the number of samples and 100 the number of features.

3.2. Scaling

All the features did not have the same scale. Therefore, two scales were dependent upon the feature selection procedures - Gaussian Scaling and Normalization. Negative values in NLP play key roles as it specifies the meaning of the same work in different contexts, hence, in order to preserve the meaning of the data, we normalized the data between (-1,1).

4. Methodology

The pipeline is visualized in the Fig 1. After feature extraction, the data was split into training and testing components, with testing comprising of 30% of the original data, and validation comprising of 20% of the remaining training data. Since there was imbalance in the two classes, we used stratified sampling. Since class imbalance was present, with the positive samples in minority, we divided training data sets into two different parts, one without any changes, and the having synthetic samples to compensate for the imbalance using SMOTE. The training data is scaled for two different feature selection procedures. Since principal component analysis (PCA) works best on standardized data, we used standardized scaling for PCA based learning. For other feature selection such as Anova, we used normalized data. The same transformation is later applied on validation and test data for evaluation and testing purposes. After applying PCA and Anova, various models were tried and cross validated with the validation data. As of now, the testing data remains untouched, as it will only be tested by the by model, which we shall select after further analysis of standard classification metrics like F1 score and ROC curves on both training and validation data, and containing phenomenon like over-fitting and under-fitting thoroughly. Till now, two models have been trained, support vector machines and logistic regression, on specific hyper-parameters.

4.1. Logistic Regression

Learning Rate :- 0.01
Optimizer :- Stochastic Gradient Descent
L2 norm was used for regularization

4.2. Support Vector Machine

Learning Rate :- 0.05
Regularization Parameter C :- 1
Kernel :- RBF
degree of the kernel :- 3 Kernel Coefficient :- 'scale'

5. Results and Analysis

Model	PCA	Anova
Logistic Regression	0.80	0.80
SVM	0.79	0.74

Table 1 : Model Comparisons between PCA and Anova (AUC scores on validation set)

5.1. Verificiation for Synthetic Data

First, we investigated if the SMOTE algorithm is able to learn the underlying distribution well enough to generate synthetic data. The figure shows t-SNE plots for the two training sets - one without SMOTE and one with SMOTE. This suggests that SMOTE worked well and can be used to deal with the class imbalance problem.

5.2. Feature Selection and Training

Performing exploratory data analysis is not reliable for high dimensional data. Therefore, inorder to analyse feature properties and the complexity of the problem, we performed two feature selection algorithms - PCA[] and Anova[]. PCA is a very robust and fast algorithm, which is still used as a state of the art method. It generates principal components, which are linear combination of original features, and they can be sorted to only select the first few according to how much information is contained in them. When we performed PCA on the standardized data, we found that first three components had covered 95% of the information. However, when we the biplots between first and second component, the classes were not at all separable. Infact, it has formed three clusters, with both the classes contained in all of them. This can be due to the fact that PCA performs linear transformation to capture maximum information, hence it may have lost information in case the data is not linear. To test this we trained both Logistic and SVM models on 3 components, and gradually increased the components to 90. As we had expected, the higher number of components gave better results on both training and validation sets. Consistently, the models performed at higher components, beating the lower components by 10-15%. It suggested that although the first few components were important, the differentiating features were present uniformly over all the components. Next, we tried Anova, which removes features which have extremely low correlation with the target variables. This method selected 95 features out of 100.

5.3. Performance measure

We generated confusion matrix, F1 score, accuracy and AUC curves for both the models. As shown in table 1, PCA gives better results for SVM, but for logistic regression, the performance is approximately same. It has been shown that for regression problems, PCA gives worse results, as in discussed in the forum [13]. Since Logistic Regression are linear in nature, it verifies the claim, as the results for both Anova and PCA are similar. Further, in this problem dimensionality reduction doesn't seem to be working, suggesting that all the features generated have some percentage of differentiating features for interacting or not. Further, both the models seem to be generalising well, considering the close values of training and validation sets. Further it suggests that the decision boundaries are linear in nature, which was not identifiable by t-SNE or PCA biplots. It needs to be further investigated. All other scores are given as a .txt file for both training and validation sets in the supplementary. AUC curves are also provided in the supplementary.

6. Conclusion

In the first deadline, we spent our time in extracting features from the raw sequence data using Doc2Vec model. Although we had done some precprocessing, in this dataset, we deeply studied the domain of the data, and realized that the earlier preprocessing may not be correctly. Hence in this deadline, we learnt :-
1) Proper way to preprocess the data - There is no single way.
2) Feature selection and dimensionality reduction, methods such as PCA and Anova. Implementing them ie easy, understanding the underlying concepts is hard. So we learnt to visualize their meanings (biplots, t-SNE).
3) SMOTE to deal with class imbalance problem.
Since we did not have much time, we were only able to train 2 models. In the next deadline, we plan to perform :-
1) Deep model analysis with algorithms specific visualizations
2) Model variance and bias calculations.
3) Selection of a model (based on cross validation) to test on the testing data.
4) Multilayer perceptron and its interpretability by studying its each layer. 5) Grid search for proper hyperparameter tuning.

7. References

[1] Lazebnik Y. Can a biologist fix a radio? *Cell*, 2002. [DOI : 10.1016/s1535-6108(02)00133-2]

[2] Ravindran V. et al. Network controllability analysis of intracellular signalling reveals viruses are actively controlling molecular systems. *Nature*, 2019

[3] Brito Anderson F. et al Protein–Protein Interactions in Virus–Host Systems, *Frontiers* in Microbiology, 2017.

[4] Yang et al. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Elsevier*, 2020.

[5] Dey et al. Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins. *Biomedical Journal*, 2020.

[6] Zhou et al. A generalized approach to predicting protein-protein interactions between virus and host, *BMC Genomics*, 2017.

[7] Barman et al. Prediction of Interactions between Viral and Host Proteins Using Supervised Machine Learning Methods, *PLOS ONE*, 2014 [DOI : https://doi.org/10.1371/journal.pone.0112034]

[8] Asgari et al. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics, *PLOS ONE*, 2015.

[9] Alguwaizani et al. Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids, *Journal of Healthcare engineering*, 2018.

[10] Haldar et al. Review of computational methods for virus–host protein interaction prediction: a case study on novel Ebola–human interactions, *Briefings in Functional Genomics*, 2017.

[11] Messina et al. COVID-19: viral–host interactome analyzed by network based-approach model to study pathogenesis of SARS-CoV-2 infection, *Journal of Translational Medicine*, 2020.

[12] https://thebiogrid.org/project/3

[13] https://www.researchgate.net/post/Is_there_a_specific_reason_that_using_PCA_gives_worse_results_than_without_using_it_in_SVM_classification