Identifying pathogen-interacting proteins in Pubmed abstracts

Rahul Natarajan[1], David Enard[4], Nandita Garud[2,3]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA, 2 Department of Human Genetics, David Geffen School of Medicine, UCLA, 3 Department of Ecology and Evolutionary Biology, UCLA, 4 Department of Ecology and Evolutionary Biology, University of Arizona

ABSTRACT

Recently more than two thousand human virus interacting proteins (VIPs) and Plasmodium (malaria)- or Piroplasm- interacting proteins (PPIPs) were discovered by manually scanning the literature and/or performing experimental procedures. Viruses and malaria have been shown to drive positive selection in humans, indicative of the burden that pathogens place on humans. However manual literature reviews are unscalable for discovery of additional VIPs, PPIPs, or novel classes of interacting proteins, e.g. bacteria interacting proteins (BIPs). To address this, here we introduce a natural language processing approach trained on abstracts from the NCBI PubMed database to automatically detect abstracts containing evidence for interacting proteins. Our results indicate, due to high AUROC and AUPRC performance on various datasets, that the model can accurately identify novel interactions between human proteins and pathogens. We apply this model to identify novel VIPs as well as 164 BIPs.

BACKGROUND AND MOTIVATION

Pathogens have been observed to impose a strong selective pressure on humans. In invading a host, pathogens often physically interact with host cells - specifically proteins - in an attempt to multiply. In response, human proteins evolve to combat these pathogens, meaning that pathogens are one of the strongest drivers of positive selection in humans [1]. Understanding the way in which pathogens and human proteins coevolve can better help an understanding of ways in which the human body combats infection and disease.
Because of the importance of pathogen-protein coevolution in terms of disease, a high throughput high accuracy methodology is necessary to identify and annotate all pathogen-protein physical interactions. Current methodologies of doing so include literature scans, mass spectrometry, and alternative experimental procedures. Mass spectrometry is a higher throughput procedure but carries a great risk of false positives. In terms of the literature approach, recently 100,000 abstracts were manually curated and manually scanned to find 1,920 virus-interacting proteins (VIPs) and 490 plasmodium-interacting proteins (PPIPs) [1]. These proteins were experimentally validated to have high rates of adaptation. However, manually scanning the literature for mentions of pathogen-protein interactions is potentially a low throughput procedure - for virus-protein interactions, it took years to find these such interactions.
Thus, we propose a high throughput high accuracy method of automated literature scanning using natural language processing (NLP) to identify new VIPs, PPIPs, and novel pathogen-interacting proteins such as bacteria-interacting proteins (BIPs). We demonstrate that our model has the ability to accurately and with high throughput predict if an abstract in Pubmed contains mention of a pathogen-interacting protein.

METHODS

From the previous literature scans, we had a set of 1,920 VIPs and 490 PPIPs [1]. Each of these had a corresponding Pubmed abstract ID from which they were obtained. To identify abstracts relevant for identifying VIPs, we queried the Pubmed database with the query terms “virus” and “human”, paginating our results. Each of the 1,920 true positives were downloaded as true positives; all other abstracts were considered true negatives. Similarly, for the malaria dataset, we queried the Pubmed database with the terms “malaria” and “plasmodium” and “gene”, again paginating. We downloaded abstracts corresponding to true positives and true negatives that were previously identified. For the novel bacteria dataset, as we had no true positives, we queried the Pubmed database with the terms “bacteria” and various HGNC symbols while excluding the terms “virus” and “mice” to ensure a lack of intersection between the datasets.

TF-IDF Tokenization
We tokenized our abstracts for training in two ways for model comparison. The first method we used was n-gram TF-IDF tokenization. For each dataset, we preprocessed the abstract text data by converting everything to lowercase letters and removing stopwords Then, we converting the abstracts into n-grams. After converting the abstracts to series of n-grams, we vectorized our input using the term frequency - inverse document frequency formula (TF-IDF).

Multilayered Perceptron (MLP)
We used a natural language processing (NLP) algorithm to discover new VIPs, PPIPs, and BIPs for the TD-IDF tokenization method. The NLP algorithm, which is a multi-layered perceptron, was built off of the algorithm described in [3]. The model has an input layer that accepts TF-IDF n-gram scores from tokenized abstracts, a hidden layer with a ReLU activation function, and an output layer that classifies abstracts as containing evidence for an interaction with a human protein via a sigmoid activation function.

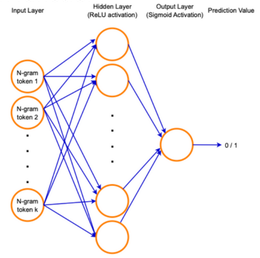

Figure 1 Model architecture of multilayered perceptron. ReLU(x) = max(0,x) and Sigmoid(x) = 1/(1+e^(-x))

BlueBERT Embeddings
An alternative tokenization technique we considered was using the embeddings generated by the large-language model BlueBERT [2]. BlueBERT is a transformer-based model, meaning that it creates a geometric representation of a word and its surrounding context. BlueBERT is a pretrained transformer-based model trained on the context of Pubmed scientific papers.

Logistic regression
We used logistic regression to discover new VIPs, PPIPs, and BIPs for the embedding tokenization method. Logistic regression classified with a weighted sigmoid.
We trained each dataset with an approximately balanced negative and positive dataset in 60/20/20 train/validation/test splits.. For both the MLP and logistic regression, we applied cross-validation to perform hyperparameter tuning for the various model parameters.

Bacteria set curation
In order to curate a set of true positives to obtain a BIP dataset, we applied the VIP-trained model to bacteria abstracts and performed manual scoring to determine the true positives.

New VIP set scoring
We obtained a list of mass-spectrometry-found VIPs. For each HGNC symbol corresponding to each of these VIPs, we downloaded abstracts by querying Pubmed with the terms “virus”, “human” and the HGNC symbol. We called this set our "recapture" set. We then applied our VIP model to this new list. For all HGNC symbols not in the original list and the "recapture" list, we repeated the above procedure and called this set the "negative" set

RESULTS

SHAP is a tool to identify important features in machine learning networks, working by iteratively choosing features to "drop" and assessing that effect on the model.

SHAP values from model indicate that expected words have high importance in MLP model

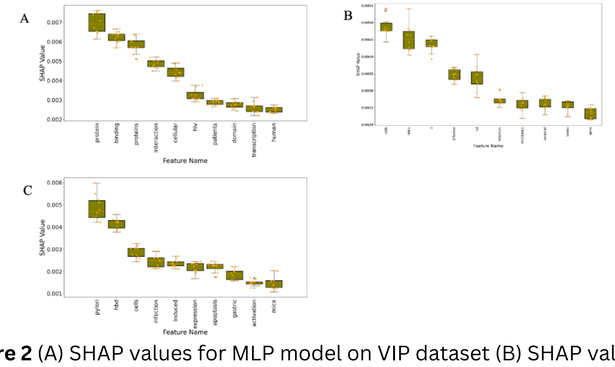

Figure 2 (A) SHAP values for MLP model on VIP dataset (B) SHAP values for MLP model on PPIP dataset (C) SHAP values for MLP model on BIPs

For each dataset, the SHAP values were calculated over 10 trials. The SHAP values for each dataset were expected. In **Figure 2** we can see that the SHAP values contain terms we would expect to see such as “interaction” and “binding”. This indicates that the model is treating the features that we would expect to be important as important

AUROC of models is high, indicating good model performance

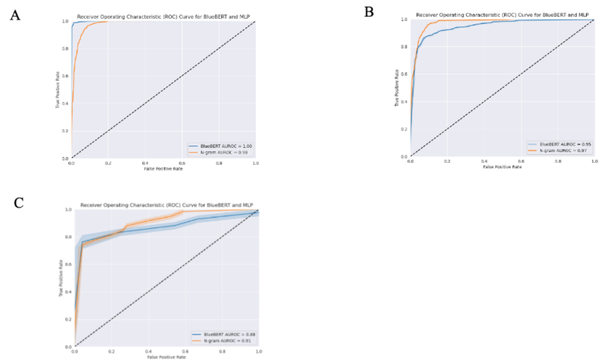

Figure 3 (A) AUROC and ROC curve for BERT (blue) and MLP (orange) model on VIP dataset (B) AUROC and ROC curve for BERT (blue) and MLP (orange) model on PPIP dataset (C) AUROC and ROC curve for BERT (blue) and MLP (orange) model on BIPs

From **Figure 3,** we see that both the BlueBERT regression model and the MLP model are highly performant. The ROC curve measures false positive rate and true positive rate at various thresholds. In both the VIP and PPIP case, we were able to obtain more training data, the BlueBERT model outperforms the MLP model in a comparison of AUROC and AUPRC.

Model scores on new abstracts falls under expected distribution

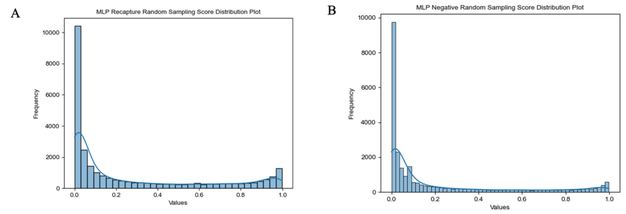

Figure 4 (A) MLP model score distribution for "recapture" set (B) MLP model score distribution for "negative" set as control

When we apply our model to the "recapture" set, we expect to see a bimodal distribution of scores where we see a high volume at both 0 and 1. We expect to see a number of positives. From **Figure 4A,** we see that the model does indeed fall under a bimodal distribution with peaks at 0 and 1. We apply to model to the "negative" set in **Figure 4B** as a control. In **Figure 4B**, we see the expected smaller peak at 1, corresponding to the discovery of novel VIPs.

IMPACT

The high AUROC and AUPRC, as well as the scoring distribution, indicate that the MLP model can indeed be used to curate a high throughput list of abstracts containing mention of interactions between pathogens and proteins. This would facilitate an easier discovery of human proteins on which adaptation to disease could be discovered.

Acknowledgements: Thanks to the BIG Summer program and Garud Lab for facilitating this project.

REFERENCES

1. **D**avid Enard, Le Cai, Carina Gwennap, Dmitri A Petrov (2016) Viruses are a dominant driver of protein adaptation in mammals eLife 5:e12469 Your paragraph text
2. Ashish Vaswani et. al (2017) Attention is All You Need https://doi.org/10.48550/arXiv.1706.03762
3. “Introduction | Machine Learning | Google for Developers.” Google, developers.google.com/machine-learning/guides/text-classification?hl=fr. Accessed 8 Aug. 2023.