Machine Learning Techniques for Sequence-Based Prediction of Viral-Host Interactions

Sikender Mohsienuddin Mohammad[#1] DilipKumar Varma Jetty[#2]

[#1]*Vice President at MUFG Union Bank & Department of Information Technology*

[#2]*Vice President at MUFG Union Bank & Department of Information Technology*

Abstract

In predicting the PPIs between the virus and human proteins, different machine learning models have been developed that are further verified using biological trials. Their classification techniques are consistent with the predictions of various sequence-based human protein properties, such as the structure of amino acids, the structure of pseudo amino acids, and correlational triads. This paper will focus on the SARS-CoV-2 virus, hepatitis E virus-human, and hepatitis B virus-human PPIs to illustrate how machine learning techniques are used to predict a sequence-based viral to host interaction. This research can promote the detection of possible targets for the more efficient production of anti-viral drugs, which have now affected the entire world. The basic well-known automated machine learning methods are widely implemented to predict PPIs, such as Random Forest, Naïve Bayes, and SVM, are used to determine the output measure focused on five-fold methods of cross-validations. In the last phase of viral infections in the host, protein-protein interactions (PPIs) play an important role. Human cells, however, are made up of a vast number of proteins. Consequently, it is just not possible to verify all different combinations of interactions through laboratory experiments. It contributes to the emergence of different computational techniques to predict and further verify PPIs in between virus and human proteins through biological experimentation. For drug usage, an understanding of how the PPIs virus proteins associate with cells of the host for reproduction and survival is important. One way that the virus associates with its host is Protein-Protein Inter-action (PPI). Therefore, predicting the PPIs between both the host and viral proteins helps in explaining how well the disease replicates and induces those virus proteins. This paper will illustrate how machine learning is used in predicting the sequence of a viral to host interaction by considering the case of SARS-CoV-2 virus, hepatitis E virus-human, and hepatitis B virus-human PPIs virus.

Key Words

Machine Learning, Supervised classification, SARS-CoV-2 virus, hepatitis E virus-human, hepatitis B virus-human, Protein-protein interaction.

Introduction

Presently, to predict PPIs, several prediction models that are based on machine learning have indeed been suggested. The interaction of the virus-host PPIs plays an important aspect in pathogenesis, as it determines

viral-host infection and host protein control. For therapeutics, the detection of main viral-host PPIs have the most significant implication (Dey et al., 2020). By incorporating various features, such as domain-domain interaction, sequence information, and network topology, utilizing VirusMINT viral-host PPIs, a comprehensive effort has been made towards anticipating viral-host PPIs. The three possibly the best-controlled machine learning techniques are widely used during the prediction of PPIs, including Random Forest, Naïve Bayes, and SVM, to test the output measure based on five-fold approaches of cross-validation. In order to determine large-scale inter-species viral-human PPIs, these machine learning techniques can be implemented. The suggested techniques can predict large-scale inter-species viral-host PPIs. This can be done through optimized machine learning models, the nature and role of the unidentified viral proteins HEV and HBV, whereby the host protein's interacting partners can be identified. Several computational techniques to predict PPIs have been created till now, and most of the ML-techniques predict PPIs inside a single species (Zheng et al., 2019). They could not predict PPIs among various species since they do not differentiate between protein interactions of species of the same type and protein interactions of species of different types. Very few machine-based techniques that are used in predicting virus-host PPIs utilizing ML methods have been developed recently.

The implementations of ML and systems biology in Sequence-Based Prediction of Viral-Host Interactions vaccine design and production are represented in a new information weight. Combining the two methods would vastly improve healthcare, speed up clinical trial procedures and reduce the expense and time spent testing and producing drugs. This paper discusses the fundamentals of approaches to machine learning and systems biology in the production pipeline for Sequence-Based Prediction of Viral-Host Interactions vaccines (Young et al., 2020). Expert Opinion machine learning and biology of systems provide the ability to escape the traditional process for vaccine production deficiencies and limitations. The implementation of both techniques in a parametric view through an effective pipeline is one promising approach. The world is entering an 'in silico age' in which research collaborations are important to resolving the long and risky path of vaccine discovery and production, along with the growing formation of the ecosystem of partners and various approaches. Regulatory guidelines should be established in this framework to qualify in silico trials as proof for the development of advanced vaccines.

Researchers, as well as in-depth studies, have assessed the issue of recognizing PPIs in recent times. Several attempts to address this issue have emerged, over and over again. While machine learning techniques are commonly used in predicting PPIs, there is still a lack of predictor variables that can draw predictions efficiently and accurately (Sudhakar et al., 2020). Many protein-protein interactions (PPIs) between both the virus and the intended host are subjected to the viral infection. Such interactions vary from either the previous attachment of the proteins' viral coat to

a membrane receptor of the host to virus proteins attacking the host's machinery of transcription. Through contamination with pathogenic viruses, different viral infections are spread. For example, the Ebola virus which is an extremely contagious and deadly illness which is caused by the Ebola virus's infection (Halder et al., 2018). Currently, there is no unique vaccination or successful treatment for this virus disease available since then. The virus-induced method is not understood fully, given the considerable number of established virus-host PPIs. Thus, machine learning models help in understanding the process of viral infection and improve therapies and vaccines by detecting interactions amongst virus proteins and host proteins.

Literature Review

Machine Learning Techniques for Sequence-Based Prediction of Viral-Host Interactions

In the previous years, experiments for PPI identification techniques have been generated. Although PPIs can be individually determined using different biophysical, biochemical, and genetics, high-throughput experimental techniques like mass spectroscopy (MS) and two-hybrid yeast (Y2H) have made it easy to determine large-scale PPIs. It includes those that have already been commonly used to predict protein functions and explain the subsequent biological processes (Kannan et al., 2020). Nevertheless, such experiments that is coupled with high-throughput displays are mostly used when classifying PPIs in a situation where the inter-species interactions have stayed underexplored. Furthermore, it is usually cumbersome, time consuming, and hard to acquire an entire protein interaction for PPIs' experimental determination. Therefore, by offering experimentally predictive models, successful computational techniques of PPI prediction will complement new methods and remove protein pairs that have a low likelihood of restricting with the PPI candidate spectrum (Wardeh et al., 2020). While primarily concentrating on predicting intraspecies PPIs, ML-driven techniques to PPI prediction are progressively used to assess interspecies PPIs like human and viral protein interactions. Some techniques accommodate for residue physicochemical characteristics of protein sequences by encoding protein sequence data and neglect the relationships between fragments of amino acids as just a component of that entire protein sequence background (Cook & Jensen, 2018). Besides, almost all models that are there are designed for some specific virus species, restricting their generalization to many other models of host-virus interactions. Significant numbers of viral-host PPIs have currently become established experimentally, offering an unparalleled quantity of information to establish generally applicable ML-based techniques for predicting human protein interactions with any virus.

When creating an ML technique for viral-host PPI prediction, the main process is to perform encoding of the function. It helps in converting viral and human protein sequences to a vector that has a fixed dimension. A few other common sequences encoding systems, including Local Descriptor, Auto Covariance (AC), and Conjoint Triad (CT), are commonly used for PPI

prediction. Whereby residue-specific physical and chemical characteristics or effects of interaction have to some degree be considered (Brito & Pinney, 2017). However, such manually constructed selected features have two drawbacks. The first is that semantic knowledge, including residues' order in whole sequences, is generally not adequately considered by such approaches. Another is that potential information from the vast amount of unidentifiable protein sequences is overlooked, although these data may reflect very significant protein characteristics.

Random Forest Technique

Random forest is a versatile and simple technique that implement ML algorithms that generates, almost all of the time, a great result even without hyper-parameter configuration. Also, it is because of its diversity and simplicity that makes it become one of the most used algorithms in the Sequence-Based Prediction of Viral-Host Interactions (Du et al., 2020). Random forest is indeed a supervised learning type of ML algorithm. The "forest" it develops resemble the decision trees structure, typically practiced in the process of "bagging," so that it can be fully operational. The main concept behind bagging involves a mixture of learning models improves even the cumulative outcome. Random forest constructs several decision trees in simplistic words and merges them to attain a more stable and accurate prediction. The main benefit of random forests is that they can be implemented during regression and classification situations. This encompasses the majority of all the existing ML systems techniques. The key drawback of random forests is that, for Sequence-Based Prediction of Viral-Host Interactions, many trees will cause the algorithm to be too weak and inefficient (Wang et al., 2020). Such algorithms are usually easy to learn, but they are very slow to make predictions once they are learned. Multiple trees are needed for a more precise Sequence-Based Prediction of Viral-Host Interactions, which leads to a slow model. The random forest algorithm is quick in most practical uses, but there will be situations where run-time efficiency is essential, and other methods are preferred.

Naïve Bayes Technique

This is basically supervised type of learning algorithm focused entirely on “Bayes theory” and is applied in solving tasks related to classification. It has been used primarily in classification of text, which necessitates a set of data that is of high-dimensional learning. The easiest and most extraordinary classification techniques is the Naïve Bayes Supervised learning model. It is because, it significantly helps in creating a very fast ML algorithm that can facilitate effective predictions. This is indeed a probabilistic classification technique, that assumes, on the assumption of an entity's likelihood in Sequence-Based Prediction of Viral-Host Interactions (Shatnawi, 2017). Also, it is called Naïve since it believes that a whole concept's incidence is independent of other features' incidence. It has been used with background experience to identify the probability of a result. The conditional probability relies on that as well. Naïve Bayes classification methods are infinitely performant. Within a learning problem, it requires a range of linear parameters in the number of attributes of the variables of a

Sequence-Based Prediction of Viral-Host Interactions (Abdelkareem et al., 2020). Instead of costly adaptive estimation as used with many other forms of classifiers, optimum training can be done by evaluating a closed-form expression that involves linear time in Sequence-Based Prediction of Viral-Host Interactions. Each characteristic makes an individual an equivalent contribution to the result. Provided the likelihood of another occurrence that has already happened, Bayes' Theory finds establishes the probability of another event taking place again.

Support Vector Machine (SVM) Technique

Support Vector Machines (SVM) is one of the most effective supervised type of learning algorithm focused mostly on hyperplane principle and is a generalized form of the maximum margin classifier, which is an intuitive and simple classifier. The algorithm produces an optimal hyper-plane for just a specified training dataset that maximizes the distance around various groups' datasets. Each category's closest objects must be well removed from either the decision boundary for a two-class dataset. SVM involves a maximal margin classifier's finding because some of the training observations may contradict that (Zhou et al., 2018). SVM is an algorithm for supervised machine learning that is used for either regression or classification problems. Although, it is commonly applied in classification issues. A plot of each piece of data within the SVM algorithm is developed to establish a position in n-dimensional range in which n represents the number of characteristics available, with each character's value being the original value of a certain coordinate. Furthermore, classification is performed through the hyper-plane discovery, which distinguishes the two groups very well (Ma et al., 2020). The technique then can categorize new text after providing an SVM model with a collection of named datasets for each group. How SVM operates can be better understood using a simplified example, which presents Support Vector Machines' fundamentals. It can be assumed that there are two tags: blue and red, and there are two features for the statistics: x and y. The main goal is to achieve a classifier that performs either blue or red, provided a combination of (x, y) points. An SVM considers these points and outputs as the hyperplane, which best separates any label, essentially a line in two dimensions (Kim et al., 2017). The line is the limit of the ruling: identifying anything that lies to one part of it as blue and anything else that lies to another as red. Using this technique, the SVM model can perform a Sequence-Based Prediction of Viral-Host Interactions.

How this Research Will help the United States

In this research, several characteristics and computational techniques that are sequence-based are used to predict the future human goals of the Sequence-Based Prediction of Viral-Host Interactions, which can significantly help the United States to predict the interactions between the viral-host protein pairs. The paper also illustrates how viral-host PPIs are predicted using machine learning methods that are well supervised; Random Forest, Naïve Bayes, and SVM. The techniques implement biological information that is significantly diversified like

amino-acids and degree, disorder origins, protein-protein association, and sequence of viral-host proteins (Eid, 2017). It is also clear that specific features can predict the viral-host PPIs and at the highest degree of exactness that is distinguishable from the other models of prediction for the viral-host PPIs. Also, from the research, it is clear that the composition of viral protein amino acids partakes a crucial role in viral-host PPIs. The research shows that the SVM technique, a machine learning-based model, is regarded as the most accurate and reasonable method of predicting the unknown interaction between the viral-host protein pairs. This research can also help biologists identify potential interactions amongst new viruses and proteins and promote the creation of anti-viral drugs. In order to comprehend the dynamics and functions of microbial communities, the review of the virus-host infectious interaction is significant (Kannan et al., 2020). Fractionated and cellular viral metagenomic data produce a substantial majority of viral connections with missing host information. While relatively simple techniques have been suggested to research viral-host interactions based on the similarity between the word frequency sequences of viruses and microbial hosts, the issue is considerably underexplored. Therefore, it was proposed that ML techniques could be effectively and efficiently used in researching the viral-host infectious interactions.

Conclusion

Currently, the use of machine learning techniques is already growing for future developments, and has become a leading research technique to build deep learning models to anticipate PPIs intra-species interactions. Consequently, other characteristics like protein contextual features and host PPI structure often take part in an essential role of identifying viral-host PPIs. More efficient computational mechanisms to drive Viral-host PPI prediction towards the advanced level should be created by thoroughly accounting for such technological advances such as machine learning techniques. Several PPI prediction computational methodologies have been proposed, historically using enterology mapping and inference focused on domain-domain/motif interactions. Although Protein 3D configurations and genetic co-expression associations are also used in predicting PPIs, other than sequence data, the protein expression and structures of data for protein pair quest are normally difficult to obtain. ML-based techniques are intensively utilized to simulate PPIs with the technological development of ML and the abundance of proven PPIs. Recently, ML-based techniques train a classification algorithm to separate pairs of protein that interacts and those that do not interact from samples of query employing established PPIs. While different heterogeneous data or facts can be combined as attributes to provide a predictive structure, protein sequence data is used for most ML-based techniques. Therefore, the study of the viral-host interaction is extremely valuable, contributing to comprehensive efforts to classify the aspects in which viruses invade, derail, and then use host features to conduct their entire life activities. It is because PPIs act as a basis of cell

contact amongst the viruses and the hosts and within the complicated viral-host interaction system. It also plays a critical role in studying viral infections and the response generated by the host's immune system.

References

[1] Abdelkareem, A. O., Khalil, M. I., Elbehery, A. H., & Abbas, H. M. (2020). Viral Sequence Identification in Metagenomes using Natural Language Processing Techniques. *bioRxiv*.

[2] Brito, A. F., & Pinney, J. W. (2017). Protein–protein interactions in virus–host systems. *Frontiers in microbiology*, *8*, 1557.

[3] Cook, H. V., & Jensen, L. J. (2018). An integrative approach to virus–host protein–protein interactions. In *Computational Cell Biology* (pp. 175-196). Humana Press, New York, NY.

[4] Dey, L., Chakraborty, S., & Mukhopadhyay, A. (2020). Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins. *Biomedical journal*, *43*(5), 438-450.

[5] Du, H., Chen, F., Liu, H., & Hong, P. (2020). Network-based Virus-Host Interaction Prediction with Application to SARS-CoV-2. *bioRxiv*.

[6] Eid, F. E. S. (2017). *Predicting the Interactions of Viral and Human Proteins* (Doctoral dissertation, Virginia Tech).

[7] Halder, A. K., Dutta, P., Kundu, M., Basu, S., & Nasipuri, M. (2018). Review of computational methods for virus–host protein interaction prediction: a case study on novel Ebola–human interactions. *Briefings in functional genomics*, *17*(6), 381-391.

[8] Kannan, S., Subbaram, K., Ali, S., & Kannan, H. (2020). The role of artificial intelligence and machine learning techniques: Race for covid-19 vaccine. *Archives of Clinical Infectious Diseases*, *15*(2).

[9] Kim, B., Alguwaizani, S., Zhou, X., Huang, D. S., Park, B., & Han, K. (2017). An improved method for predicting interactions between virus and human proteins. *Journal of bioinformatics and computational biology*, *15*(01), 1650024.

[10] Ma, Y., He, T., & Tan, Y. T. (2020). Seq-BEL: Sequence-based Ensemble Learning for Predicting Virus-human Protein-protein Interaction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

[11] Shatnawi, M. (2017, August). Protein-Protein Interaction Prediction: Recent Advances. In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)* (pp. 69-73). IEEE.

[12] Sudhakar, P., Machiels, K., & Vermeire, S. (2020). Computational Biology and Machine Learning Approaches to Study Mechanistic Microbiomehost Interactions.

[13] Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J. C., Fuhrman, J. A., ... & Ahlgren, N. A. (2020). A network-based integrated framework for predicting virus–prokaryote interactions. *NAR genomics and bioinformatics*, *2*(2), lqaa044.

[14] Wardeh, M., Blagrove, M. S., Sharkey, K. J., & Baylis, M. (2020). Divide and conquer: machine-learning integrates mammalian, viral, and network traits to predict unknown virus-mammal associations. *BioRxiv*.

[15] Young, F., Rogers, S., & Robertson, D. L. (2020). Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLoS computational biology*, *16*(5), e1007894.

[16] Zheng, N., Wang, K., Zhan, W., & Deng, L. (2019). Targeting virus-host protein interactions: Feature extraction and machine learning approaches. *Current drug metabolism*, *20*(3), 177-184.

[17] Zhou, X., Park, B., Choi, D., & Han, K. (2018). A generalized approach to predicting protein-protein interactions between virus and host. *BMC genomics*, *19*(6), 69-77.