Current Trends in Biotechnology and Pharmacy
Vol. 15 (6) 80 - 82, 2021, ISSN 0973-8916 (Print), 2230-7303 (Online)
10.5530/ctbp.2021.6.14

Transfer Learning-Based Attention Gated Siamese Network for Human and SARS-CoV-2 Protein Interactions

Nikhil Sharma[1], Shivansh Gupta[2], Manas Gupta[2], Priyadarshini[3], Satish Chandra[2]*

[1]Department of Electronics and Communication Engineering, JIIT, Noida
[2]Department of Computer Science Engineering and Information Technology, JIIT, Noida satish.chandra@jiit.ac.in
[3]Department of Biotechnology, JIIT, Noida

Abstract

For the past year SARS-CoV-2 has affected the lives of people around the globe. Therefore, research community is continuously putting in their best efforts to find a solution to curb and cure the disease. SARS-CoV-2 is a 29.9k bp long sequence genome comprising of 25 different proteins among which spike glycoprotein plays a vital role in interaction with the host cells. Hence, majority of the scientific studies were focused towards targeting the spike region for the vaccine design against the contagious virus. Thorough study of protein-protein interaction between human and virus can help us in better understanding and management of this disease. For this purpose, an Attention gated Siamese framework is utilized from which a consensus of prominent features and contextual information is taken into account to identify the influence of protein sequences. Moreover, to obtain the pattern of interacting pairs of human and SARS-CoV-2 proteins, a transfer learning-based approach is opted from the proposed network through which we obtained an accuracy of 85%. Additionally, by using this model, we identified that there were 30, 13 and 17 human proteins interacting with spike glycoprotein, nucleocapsid and membrane respectively, having predictive interaction of above 90% for each of the interactions.

Keywords Virus; SARS-CoV-2; Attention gated Siamese framework; Proteins.

Introduction

For past year SARS-CoV-2 has affected the lives of people around the globe which is responsible for COVID-19. With different form of mutations and sudden surge of the pandemic disease has drawn a lot of attention.With a lot of hard work put in by research communities many vaccine candidates are proposed in form of Covaxin designed and manufactured by Bharat Biotech along with the Indian Council of Medical Research (ICMR) providing a ray hope to entire nation. Hence, it is important to study the molecular pattern of SARS-CoV-2 infecting the cells as well as to understand how mutation in the virus will and can affect the human cells. SARS-CoV-2 virus is a single-stranded enveloped RNA virus of length 29.9 kilobase pair which is one of the largest known RNA virus [1,2,3]. This virus has 11 coding regions where ORF1ab occupies majority of the genome, whereas spike (S), envelope (E), membrane (M), nucleocapsid (N) and 6 accessory regions such as ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10 occupy rest of the genome. A major role in recognizing the host receptor is played by the Spike region[4] and the shape of the virions is influenced by Envelopeas well as Nucleocapsid plays a vital role of interferon inhibitor via packaging the genomes.

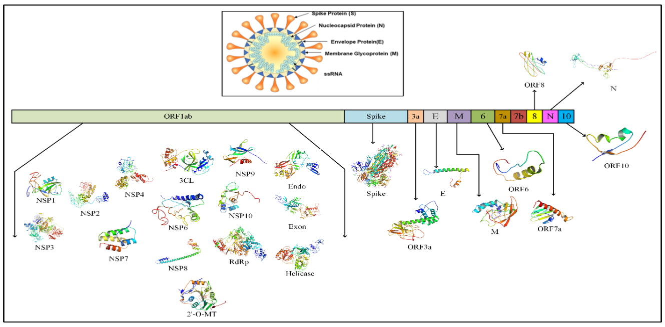

Figure 1 SARS-CoV-2 protein distribution between 29.9k long RNA genome sequence.

Protein–Protein Interaction (PPI) provides a medium to study about the interaction process of the host and virus cells which is essential for survival and replication of the foreign agent in the host environment.In Zoul et. al.[5], identified the envelope and membrane proteins of SARs-CoV-2 to be highly similar with that of SARs-CoV. Further, with the help of network proximity analyses between 16 prioritized anti-HCoV repurposable drugs. Moreover, authors in the study identified 3 potential drug combination through Complementary Exposure pattern. Hence, such efficient interaction methods can be taken into account to cover the void through testable hypothesis. Due to the high expenses of experimental approaches for identification of virus-host PPIs, Khorsand et. al utilized the effectiveness of three-layered network considering Alpha influenza virus proteins which were more similar to SARS-CoV-2 proteins, followed by the second phase in which protein-protein interaction information between Alpha Influenza and human proteins. In the final layers protein-protein interaction was studied between the SARS-CoV-2 and human proteins through a clustering coefficient network property on the later layers. In case of ongoing SARS-COV-2, gordon et. al identified 332 high-confidence protein–protein interactions between SARS-CoV-2 and human proteins using the mass spectrometry [7].In another study, by Dey et. al predicted potential human-SARS-CoV-2 interacting pairs based on the amino and pseudo amino acid composition along with the conjoint traid features with help of Learning Vector Quantization (LVQ) to reduce the features vectors and then further employed multiple machine learning techniques such as Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and K-Nearest Neighbor (KNN) for classification and prediction of the potential interacting SARS-CoV-2 and human protein pairs The above studies use the classical Machine Learning and clustering approaches to find the interactions but lacks in exploring the strength of Deep learning models. Therefore, we propose a deep learning approach for finding the interacting patterns in Human-virus protein sequences.

In this study, we have performed a transfer learning-based approach to extract the interacting pattern between human-human interacting protein pairs with the help of Siamese based Attention framework and predicted the probability of interaction between virus and human proteins.

Materials and Method:

Collection of human protein interaction data

Protein interaction data were taken from 3 sources: BioGrid[9], MINT[10] and HPRD[11] and was used to train and test the model. Due to the large size of database, a random subset of data was created. 300k pairs from the BioGrid dataset were randomly extracted and taken into account for this study.

Structural and Physio-chemical descriptors

In this study, molecular interaction pattern between human-human proteins is extracted and considered to obtain the interaction similarity in between the SARS-CoV-2 and human proteins. For which initially, Molecule Descriptors such as C/T/D, Auto-corelation, Conjoin Triad and Quasi Sequence order were extracted for each protein sequence.

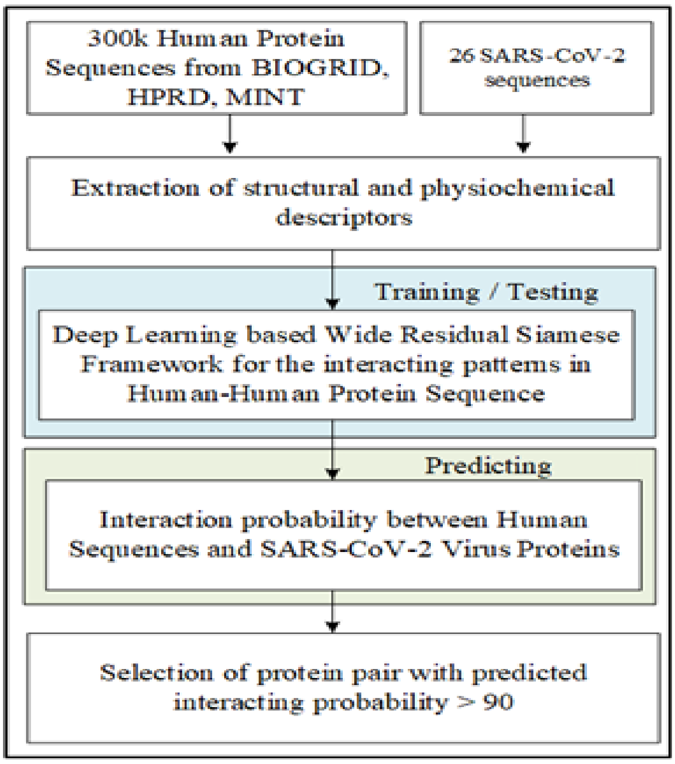
300k Human Protein Sequences from BIOGRID, HPRD, MINT
26 SARS-CoV-2 sequences
Extraction of structural and physiochemical descriptors
Training / Testing
Deep Learning based Wide Residual Siamese Framework for the interacting patterns in Human-Human Protein Sequence
Predicting
Interaction probability between Human Sequences and SARS-CoV-2 Virus Proteins
Selection of protein pair with predicted interacting probability > 90

Figure 2 Proposed Workflow for the protein-protein interaction

Siamese Network

This network was constructed using two parallel identical neural networks. Each of these parallel units consisted multiple CNN blocks[12].

Proposed Framework

Following the pre-processing steps, a matrix pool was created of multiple descriptors to represent a protein which is fed as an input to Deep Learning framework.For the same, an attention gated Siamese nework was used in which two parallel identical Convolutional Networks were taken into account to extract the prominent features and contextual information from the interacting protein sequences. Moreover, to obtain the pattern of interacting pairs of human and SARS-CoV-2 proteins, a transfer learning-based approach is opted to predict the interaction probability in between the human and virus protein pairs.

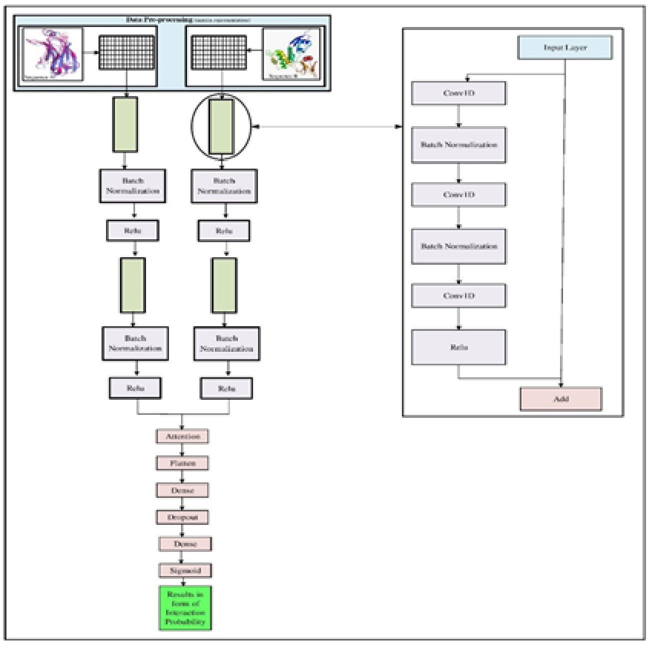

Figure 3 Siamese Attention gated framework used for Protein-Protein interaction

Results and Discussion

it is worth to note thatprediction carried out for protein-protein interaction in between SARS-CoV-2 virus and human is highly influenced by the training of proposed network on the Human-Human protein network hence, to optimize our proposed model's performance, a dense network is designed alongside with L1 and L2 norm regularizers.As a result, an architecture was achieved with accuracy of 85.97% and 81.2% AUC along with 83% sensitivity and specificity of 84% (Figure 4).

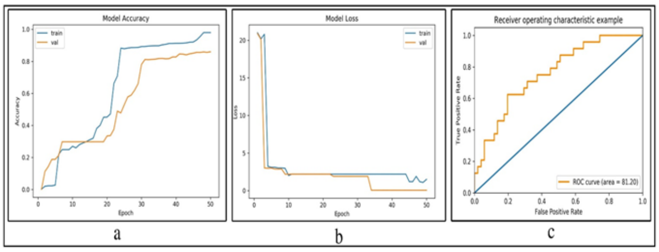

Figure 4 Model performance with help of a) accuracy b) Loss (Binary Cross Entropy) c) ROC curve

In this study, as shown in Figure 5, we report the highest interacting SARS-CoV-2 and human protein pair whose predicted interaction were above 90%. As a result, we obtained 30, 17, 16, 15, 23, 13, 15, 34, for Spike, ORF8, NSP9, NSP7, NSP12, Nucleocapsid, Membrane, Helicase respectively(Figure 6). Among these highly interacting pairs rank wise analysis was done to select the best candidate.

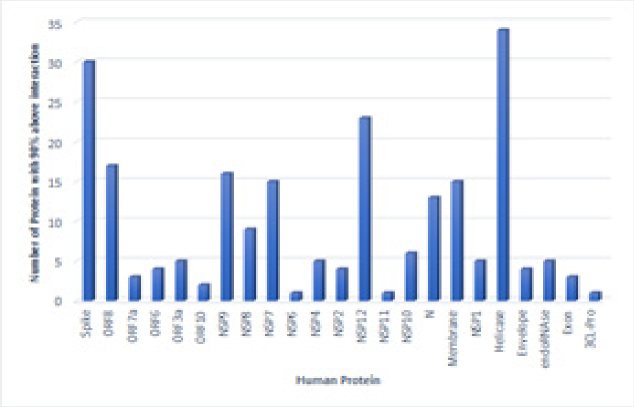

Figure 5 Number of protein pairs interacting with the 23 SARS-CoV-2 proteins

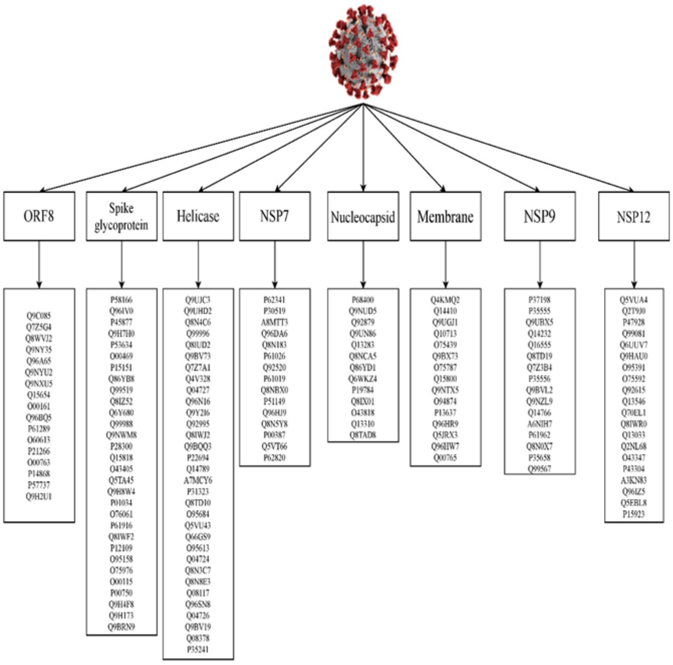

Figure 6 Interaction of human protein and SARS-CoV-2protein which are above 90%a)ORF8 (17)b) Spike (30) c)Helicase (35) d)NSP7 (15) e)Nucleocapsid (13) f)Membrane (15) g) NSP9 (16) h) NSP12 (20)

Conclusion

In this work, we have analysed300k+human-human protein pairs to extract the interaction patterns between the proteins. Further, with help of transfer learning-based approach we identified the similar interaction patterns within the SARS-CoV-2 and human protein pairs. As a result, we found significant number of virus-human protein pairs which can be considered as potential target to curb the disease carried by the contagious vrius. Hence, such evolving deep-learning approach can be helpful in drug design of such frequently mutating virus.

References

[1] Alanagreh, Lo'ai&Alzoughool, Foad&Atoum, Manar. (2020). The Human Coronavirus Disease COVID-19: Its Origin, Characteristics, and Insights into Potential Drugs and Its Mechanisms. Pathogens. 9. 331.

[2] Di Wu, Tiantian Wu, Qun Liu, Zhicong Yang, The SARS-CoV-2 outbreak: What we know, International Journal of Infectious Diseases, Volume 94,2020,Pages 44-48,ISSN 1201-9712.

[3] Andersen, K.G., Rambaut, A., Lipkin, W.I. *et al.* The proximal origin of SARS-CoV-2. *Nat Med* **26,** 450–452 (2020).

[4] Shang, Jian, et al. "Cell entry mechanisms of SARS-CoV-2." *Proceedings of the National Academy of Sciences* 117.21 (2020): 11727-11734.

[5] Zhou, Y., Hou, Y., Shen, J. *et al.* Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* **6,** 14 (2020).

[6] Khorsand, B.,A., Savadi, M., Naghibzadeh. SARS-CoV-2-human protein-protein interaction network,Informatics in Medicine Unlocked,Volume 20,(2020),100413 ISSN 2352-9148

[7] Gordon, D.E., Jang, G.M., Bouhaddou, M. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583,** 459–468 (2020).

[8] Dey, L., S., Chakraborty, A., Mukhopadhyay, Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins, Biomedical Journal, Volume 43, Issue 5,2020, Pages 438-450, ISSN 23194170. https://doi.org/10.1016/j.bj.2020.08.003

[9] Chatr-Aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. MINT: the Molecular INTeraction database. Nucleic Acids Res. 2007;35(suppl 1):572–4. 27.

[10] Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, et al. The BioGRID interaction database: 2011 update. Nucleic Acids Res. 2011;39(suppl 1): 698–704.

[11] Chatr-Aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G. MINT: the Molecular INTeraction database. Nucleic Acids Res. 2007;35(suppl 1):572–4. 27.

[12] Keshava Prasad, T. tempspacetempspaceS, et al. "Human protein reference database—2009 update." *Nucleic acids research* 37.suppl_1 (2009): D767-D772.

[13] J. Shen, X. Tang, X. Dong and L. Shao, "Visual Object Tracking by Hierarchical Attention Siamese Network," in *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3068-3080, July 2020.