Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 12 NO: 1 PAGE: 166-179 YEAR: 2023
DOI:10.17798/bitlisfen.1220301

Computational Prediction of Interactions Between SARS-CoV-2 and Human Protein Pairs by PSSM-Based Images

Zeynep Banu ÖZGER[1*], Zeynep ÇAKABAY[1]

[1] *Kahramanmaraş Sutcu Imam University, 46040, Kahramanmaras, Türkiye*

(ORCID:0000-0003-2614-3803) (ORCID: 0000-0001-7059-2337)

Keywords: Feature Projection, Pathogen-Host Protein Interaction, Siamese Neural Network, SARS-CoV-2

Abstract

Identifying protein-protein interactions is essential to predict the behavior of the virus and to design antiviral drugs against an infection. Like other viruses, SARS-CoV-2 virus must interact with a host cell in order to survive. Such interaction results in an infection in the host organism. Knowing which human protein interacts with the SARS-CoV-2 protein is an essential step in preventing viral infection. In silico approaches provide a reference for in vitro validation to protein-protein interaction studies by finding interacting protein pair candidates. The representation of proteins is one of the key steps for protein interaction network prediction. In this study, we proposed an image representation of proteins based on position-specific scoring matrices (PSSM). PSSMs are matrices that are obtained from multiple sequence alignments. In each of its cells, there is information about the probability of the occurrence of amino acids or nucleotides. PSSM matrices were handled as gray-scale images and called PSSM images. The main motivation of the study is to investigate whether these PSSM images are a suitable protein representation method. To determine adequate image size, conversion to grayscale images was performed at different sizes. SARS-CoV-2-human protein interaction network prediction based on image classification with siamese neural network and Resnet50 was performed on PSSM image datasets of different sizes. The accuracy results obtained with 200x200 size images and siamese neural network as 0.915, and with 400x400 size images and Resnet50 as 0.922 showed that PSSM images can be used for protein representation.

1. Introduction

Proteins are polymers formed by the polymerization of amino acids. Each protein has its own features due to its amino acid sequences. These sequences also determine the function of the protein [1]. Many biological events in our body occur as a result of the binding/dissociation of proteins with each other. Understanding protein-protein interactions has a critical role in drug and peptide design. Additionally, understanding the root causes of protein interactions is a big step towards controlling events at the molecular level. The proteins interact via their surface domains as shown in Figure 1. In order for it to interact, the two protein interfaces must be compatible with each other, both shapely and chemically [2].

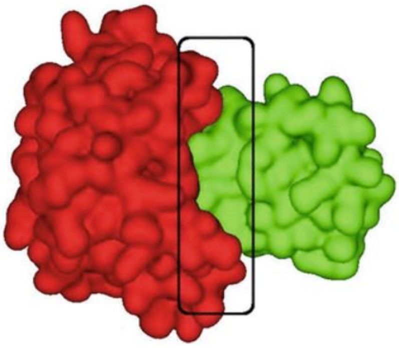

Figure 1. Protein protein interaction [8].

Detection of protein interactions is performed by in vitro, in vivo or in silico methods [3]. Interaction detection by in vitro methods is done with

[*] Corresponding author: zeynepozger@ksu.edu.tr

Received: 16.12.2022, Accepted: 26.02.2023

microorganisms, cells or biological molecules other than living organisms [4]. In vivo methods are performed on living organisms or cells, as are clinical studies in animals. It is usually applied after in vitro tests [5]. Both have time, place, and cost constraints as they must be performed in a clinical setting. In silico studies are studies performed with computers or simulations [6]. It provides an advantage in terms of reducing the need for clinical research. Considering the existence of a large number of proteins in the body, it is a costly process to investigate whether they interact with each other in the clinical setting. Instead, by identifying proteins with high interaction potential, clinical validation of only these proteins saves time [7]. In this context, different approaches are applied. Studies on the classification of medical documents [9] make it easier to reach the studies presented in a certain field. Protein primary structures [10-12], association rule mining [13], protein domain profiles [14], gene ontology [15, 16], 3D structures of proteins [17], domain-domain motif interactions [18], domain and sequence signature [19], network topology [20], phylogenetic trees [21], text mining [22] are common methods for identifying protein-protein interactions (PPI).

Virus genomes require a host cell to replicate themselves. The way of entry into the host cell is through protein interactions. That is, a protein of the virus and a protein of the host cell must interact with each other [23]. Coronaviruses are infections in which animal viruses such as bat and mink acquires the ability to infect humans [24]. In drug development for coronavirus outbreaks, targeting proteins involved in host-pathogen interactions is important to prevent the spread of epidemics [25]. Like other viruses, SARS-CoV-2 must interact with host cell proteins to reach the host cell and replicate its genome.

Although the SARS-CoV-2 epidemic seems to be under control, it continues to exist in the world. However, the presence of epidemic diseases such as MERS and SARS-CoV seen in the past is an indication that there may be various epidemics belonging to the coronavirus family in the future. Genome data from past coronavirus outbreaks has been pioneering in treatment studies developed for the SARS-CoV-2 virus. Therefore, it is important to determine the entry routes of the SARS-CoV-2 virus into the host cell, so that the world does not fall into a bottleneck, as in SARS-CoV-2, in future coronavirus outbreaks. In order to develop preventive and therapeutic drugs, researchers continue their studies to understand the interactions of SARS-CoV-2 and human proteins. So, there are a limited number of in silico studies presented in the literature. Lanchantin et al. [26], proposed a transfer learning method for predicting SARS-CoV-2-human protein interactions. The work is based on the idea that polypeptide sequences that occur between interacting proteins can be conserved in different organisms. Authors used transfer learning to learn these short protein motifs. Because known interaction data for SARS-CoV-2 is limited, the model was trained with pathogen-host interaction data of other viruses [27]. The success of the proposed method was also applied to predict the interaction of different viruses, and the obtained AUROC (area under receiver operating characteristics curve) value for SARS-CoV-2 has been reported as 0.753.

According to Du et al. [27], SARS-CoV-2 can infect humans as well as some mammals such as cats, dogs, and tigers. These infected animals can infect the virus humans. In the study, a 2-level multi-level perceptron network (MLP) with 2 levels was used to build a protein-protein interaction network. The MLP network was trained with 7 human coronaviruses and 17 hosts. In conclusion, the authors found 19 possible interactions between human and SARS-CoV-2 proteins. The SARS-CoV-2 virus belongs to beta coronavirus family. This coronavirus family has 5 subtypes: sarbecovirus, embecovirus, merbecovirus, hibecovirus and nobecovirus [28]. Since SARS-CoV and SARS-CoV-2 viruses belong to the sarbecovirus type, Khorsand et al. [29], focused on these types of viruses. The authors used a three-layer neural network in their study. The first layer contains the proteins of alpha influenza viruses similar to SARS-CoV-2 viruses. The second layer includes the known alpha influenza virus-human protein interactions. And the third one contains the known SARS-CoV-2-human protein interactions. Of the 87894 sarbevirus-human interactions found, 7201 were reported to be SARS-CoV-2-human protein-protein interactions. Khan and Khan [25] investigated protein-protein interactions for MERS, SARS-CoV, and SARS-CoV-2 and they identified common host targets with bioinformatics tools for these outbreaks. The known interactions are obtained from BioGrid [30] database. Dey et al. [31], determined SARS-CoV-2-human protein-protein interactions using machine learning techniques with sequence-based data. The algorithms were trained on 332 interactions discovered by Gordon et al. [32] using affinity purification mass spectrometry method. In the study, three sequence-based features (amino acid composition, conjoint triad, pseudo amino acid composition) were obtained from protein amino acid sequences. The decisions of SVM and random forest learners were combined with the ensemble majority voting technique. The best prediction score obtained was reported as 72.33% accuracy rate. In addition, the authors presented a gene ontology term analysis of predicted 1326 human proteins interactions. Pirolli et al. [33] aimed to find the binding receptors between the spike protein of SARS-CoV-2 and the human angiotensin converting enzyme (ACE2) protein. They focused their research on the ACE2 protein because the spike protein uses ACE2 to enter the host cell. In their study, the authors predicted binding receptors with a convolutional neural network-based quantitative structure activity relationship (QSAR) model. Lee [34] proposed a virus type-independent PPI estimation method. The authors obtained 80,775 virus-human PPI from the STRING database. They represented the PPI network of known interactions with a bipartite graph. Using the nodes in the graph, they obtained fractional compositions of 20 amino acids. Then, they extracted features from these composition profiles with 72 different distance/similarity measurements. They made predictions with models trained with random forest and XGBoost. The XGBoost model achieved the best performance with an accuracy value of 68%. SARS-CoV-2-human protein interaction prediction with the trained models was performed with an accuracy of 58%. In the study presented at [35], Bell et al. presented a pipeline they call PEPPI. PEPPI is a virus type-independent consensus model and includes structure similarity, conjoint-triad-based neural network, sequence similarity, and functional association data. The modules are combined with the naive Bayesian consensus classifier. The authors tested their model also on SARS-CoV-2-human protein interactions. The pipeline correctly predicted 94 out of 128 interactions.

A protein-protein interaction detection problem with computational methods can be handled as a binary or multi-label classification problem. The training data consists of host proteins interacting and non-interacting with a pathogen protein. There may be more than one known protein for a virus species. In the binary classification approach, it is not necessary to know which protein of the virus a host protein interacts with. That is, if a host protein interacts with any protein of the virus, its label will be 1, otherwise it will be 0. In the multi-class classification approach, however, it is necessary to know which protein of the virus interacts with the host protein. The label of each different protein of the virus is different. In the dataset, each host protein is labeled with the label of the virus protein that interacts with or does not interact with that protein of the virus.

The binary classification approach does not need to know which protein of the virus an interaction is with. It is sufficient for the virus to know that any protein interacts with a host protein. This approach is advantageous given the difficulty of obtaining validated interaction data in a lab setting. However, the trained model can only tell whether it interacts with the virus of interest for a given host protein. To train a model with a multi-class approach the model needs data with sufficient interactions for each protein of the virus. However, such a model tells not only that the given host protein interacts with the virus protein but with which protein of the virus it interacts with.

In protein-protein interaction network studies, position specific scoring matrices are generally used for inferences such as the biological information they contain and some properties of amino acids. Within the scope of this study, it is being investigated whether PSSM matrices can be used as images to realize a protein-protein interaction prediction problem for SARS-CoV-2. For this purpose, the known interacting protein pairs were converted into gray-scale images of different sizes. Based on the lack of studies on protein interaction network prediction by in silico analysis for SARS-CoV-2, a method has been proposed to identify possible proteins that could be targeted in treatment development for SARS-CoV-2. The problem was handled both as a multi-class problem using siamese neural network and a binary-class problem using Resnet50.

There are numerous approaches to extract features from proteins for protein interaction network prediction. These approaches are generally geared towards exploiting protein primary and secondary sequences and the physicochemical properties of proteins. The contribution of this study is to show that proteins can be converted into image data with PSSM matrices, and protein interaction network prediction can be made with siamese neural networks trained with positive and negative interactions.

2. Material and Method

2.1. Dataset Description

Classification algorithms need samples in each class in the dataset for training. The positive class samples are taken from the study by Gordon et al. [32]. In the dataset, there are 332 interactions between 27 SARS-CoV-2 proteins and 332 human proteins. The dataset consists of 2 columns. The first column contains the primary amino acid sequence of a SARS-CoV-2 protein and the second column contains the primary amino acid sequence of the human protein determined to interact with this SARS-CoV-2 protein. The

Table 1. Numbers of human proteins interacted with SARS-CoV-2 proteins

SARS-CoV-2 Protein	#interacting human protein	SARS-CoV-2 Protein	#interacting human protein	SARS-CoV-2 Protein	#interacting human protein
Envelope	6	**Nsp7**	32	**Orf3a**	8
Membran	30	**Nsp8**	24	**Orf3b**	1
Nucleocapside	15	**Nsp9**	16	**Orf6**	3
Spike	2	**Nsp10**	5	**Orf7a**	2
Nsp1	6	**Nsp11**	1	**Orf8**	47
Nsp2	7	**Nsp12**	20	**Orf9b**	11
Nsp4	8	**Nsp13**	40	**Orf9c**	26
Nsp5	3	**Nsp14**	3	**Orf10**	9
Nsp6	4	**Nsp15**	3		

shortest of these sequences consists of 13 amino acids and the longest consists of 5596 amino acids. Since there is more than one human protein determined to interact with a SARS-CoV-2 protein, information on how many human proteins each SARS-CoV-2 protein interacts with is given in Table 1.

In the learning phase, algorithms need both interactive and non-interactive examples. However, experiments to identify interacting protein pairs are not focused on finding non-interacting proteins. Therefore there is no gold standard for saying that a protein is non-interacting with a specific virus [36]. Different approaches such as random sampling [37], subcellular localization [38], and sequence similarity [23] techniques were used in studies to determine negative proteins. In our study, we used the sequence similarity [12] approach. The basic idea underlying this method is that the sequence similarity of host proteins interacting with a virus protein can be high. In bioinformatics, the most similar regions of different gene or protein sequences can be detected by sequence alignment methods. Thus, information such as the functions of these genes or proteins and which organism they belong to can be determined to a large extent. Substitution matrices are used to calculate the similarity scores of proteins. Substitution matrices are matrices consisting of the biological significance or randomness scores of the occurrences of the standard 20 amino acids. Blosum62 [39] was used as the substitution matrix in this process performed in the R environment.

5873 human proteins not found in the positive dataset were obtained from Uniprot [40]. These are candidate negative proteins. The sequence similarity matrix was constructed with 332 positive proteins and candidate negative proteins. The sequence identity matrix contains similarity ratios of the proteins. Rows represent the positive human proteins and columns represent candidate negative proteins. Therefore, the size of this matrix was 332x5873. Each cell includes the similarity ratio of the corresponding positive protein and the candidate negative protein.

As seen in Table 1, the number of human proteins that interact with each SARS-CoV-2 protein varies. The negative dataset was created by considering the protein counts in the positive dataset. For example, the dataset includes 6 human proteins identified to interact with the envelope protein of SARS-CoV-2 (Table 1). Of the candidate negative proteins, 6 proteins with the lowest sequence similarity to these 6 human proteins were identified. These were added to the negative dataset and labeled as human proteins non-interacting proteins with the envelope protein. This process was repeated for all SARS-CoV-2 proteins. Thus, as many non-interacting proteins were added to the dataset as the number of interacting proteins for each SARS-CoV-2 protein. As the total number of interacting proteins was 332, the number of non-interacting proteins added to the dataset was also 332.

2.2. Position Specific Scoring Matrices (PSSM)

Proteins are polypeptides formed by the covalent bonding of amino acids to each other in a certain type, in a certain number and in a certain sequence [41]. There are 20 amino acids. When amino acids come together in a different order, different proteins are formed. PSSM matrices contain the probability of occurrence of each amino acid and nucleotide at each position [42]. For proteins, the row number of the PSSM matrix is equal to the amino acid number, i.e. 20. The number of columns of the matrix is equal to the length of the protein sequence. The value of each cell is the probability that the corresponding amino acid is in the corresponding position. These probabilities are derived from multiple sequence alignment [43] scores. In bioinformatics, PSSM matrices can be used for a variety of tasks such as predicting the attributes of a protein [44]. Building a PSSM matrix is given in Figure 2.

Protein sequences were downloaded from Uniprot in Fasta format. PSSM matrices of positive, negative and SARS-CoV-2 proteins were constructed separately using Blosum62 substitution matrix in R environment with “protr” [45] package.

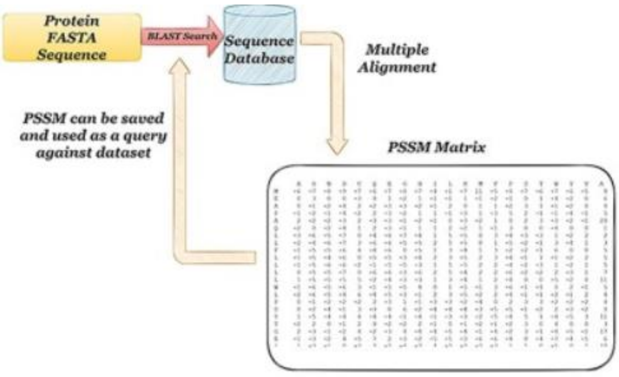

Figure 2. Building a PSSM matrix [44].

2.3. Convolutional Neural Networks (CNN)

The convolutional neural network is a deep learning architecture that is generally used for image processing. CNN uses different processes to capture features in images. Then, using these features, a CNN network classifies the images. A CNN network basically consists of a convolutional layer, a pooling layer, and a fully connected layer. Images are matrices of pixels. The purpose of the convolution layer is to try to capture certain features in the image with a filter smaller than the original image size. The pooling layer aims to reduce dimensionality [46]. In this way, computational complexity is reduced and unnecessary features are eliminated. The fully connected layer transforms the pixel matrix which passes through the convolutional and pooling layers several times into a flat vector. After these processes, images can be classified using traditional neural networks [47]. The general architecture of CNN is given in Figure 3.

2.4. Residual Neural Networks (RNN)

Resnet is an enhanced version of CNNs [49]. When deep networks begin to converge, the problem of degradation arises [50]. Resnet was developed to solve this degradation problem of CNN. In Resnet, there are shortcuts between layers. Resnet includes residual blocks to reduce training errors. The scheme of residual blocks are given in Figure 4. These shortcut links allow one or more layers to be skipped. Thus, now blocks and inputs can propagate faster over the remaining connections between layers. In this way, the degradation problem is prevented as the network deepens [49]. Resnet also uses bottleneck blocks to make training faster. The general network architecture of Resnet is given in Figure 5.

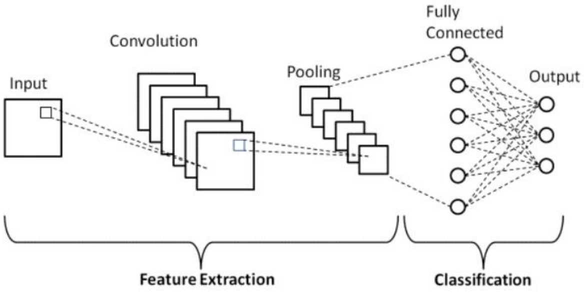

Figure 3. General architecture of CNN [48].

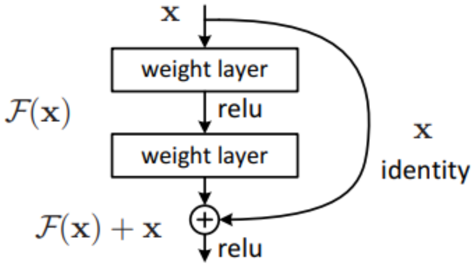
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 4. Residual blocks of Resnet [48].

Resnet50 [49] is a pre-trained 50-layer network. The authors in [49] trained the network with the ImageNet dataset. ImageNet [51] is a reference dataset created for use in object recognition research. Thus, researchers working on object recognition can perform transfer learning by applying fine-tuning on previously learned parameters according to their own datasets. In this study, we used the Resnet50 architecture for the protein interaction network prediction problems. We converted the protein sequences to PSSM images and trained the network with this images. Since protein interaction network detection is not an object recognition problem, we did not use pre-trained parameters.

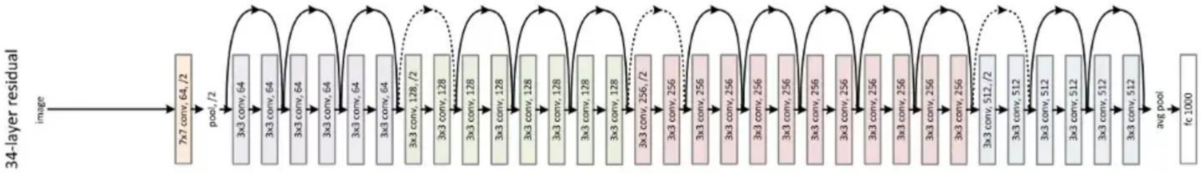

Figure 5. General architecture of Resnet [49].

2.5. Siamese Neural Networks (SNN)

The siamese network does similarity learning using two identical network architectures. Two entries are given to the network. Siamese network contains two separate neural networks. Subnets share parameters. These shared parameters allow to distinguish between two entries that are the same or different. The first subnet encodes the first input, and the second subnet encodes the second input. The Siamese network decides that these inputs belong to the same or different things, based on the difference between the two encoding outputs [52]. The general architecture of the Siamese network is given in Figure 6.

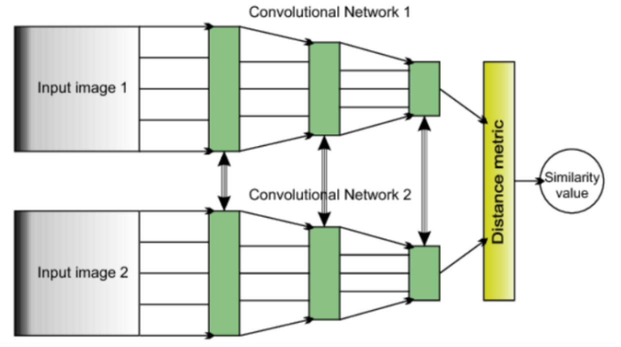

Figure 6. General architecture of SNN [53].

SNNs are powerful networks, especially for learning complex relationships between two images [54]. We trained SNN with interacting and non-interacting protein pairs. For interacting protein pairs, the first input is the PSSM image of positive protein and the second is the PSSM image of SARS-CoV-2 protein. For non-interacting protein pairs, the first input is the PSSM image of negative protein and the second is the PSSM image of SARS-CoV-2 protein. Convolutional layers learn the filters and are responsible for finding common patterns between proteins. Because the two subnets share parameters, the network is expected to find similar properties for interacting proteins. SNN takes a pair of images as input and gives the probability of similarity of these two images as output. In our problem, the image pairs supplied to the network are PSSM images of SARS-CoV-2 and human proteins with and without interaction. The output of the network is the probability whether a pair of PSSM images given to it are interactive. According to the similarity value, it is important to determine the threshold value correctly in order to decide whether the proteins are interacting or not. According to the results obtained from the positive and negative samples during the training phase, it was decided experimentally that the value of 0.5 was an appropriate threshold value (Eq. 1).

$$\boldsymbol{x_i} = \begin{cases} \boldsymbol{non-interacting, if\ similarity < 0.5} \\ \boldsymbol{interacting, if\ similarity \geq 0.5} \end{cases} \quad (\mathbf{1})$$

2.6. General Framework of the Proposed Method

In this study, it was investigated whether the use of PSSM matrices as images is suitable for the protein-protein interaction problem. The general framework for converting PSSM matrices to images is given in Figure 7. Protein sequences of human and SARS-CoV-2 proteins were obtained from Uniprot database. Uniprot [40] is a public and universal database. It contains detailed information about proteins such as protein sequences and functions. PSSM matrices were obtained using the Blosum62 substitution matrix. The size of the PSSM matrix is 20xL. 20 indicates the unique amino acid number and L indicates the length of the protein. All matrices were converted to grayscale images of different sizes (20x20, 50x50, 100x100, 200x200, 400x400). The prediction phase was applied separately for each image scale.

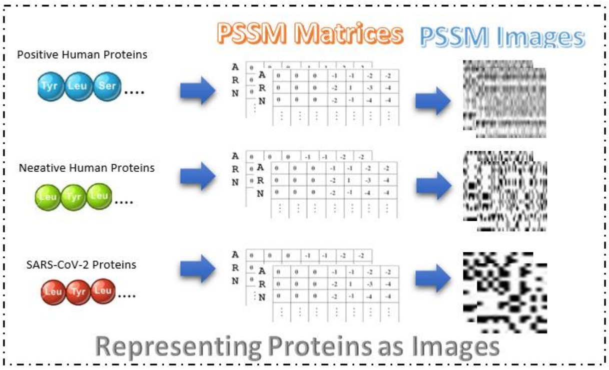

Figure 7. Encoding proteins as Images.

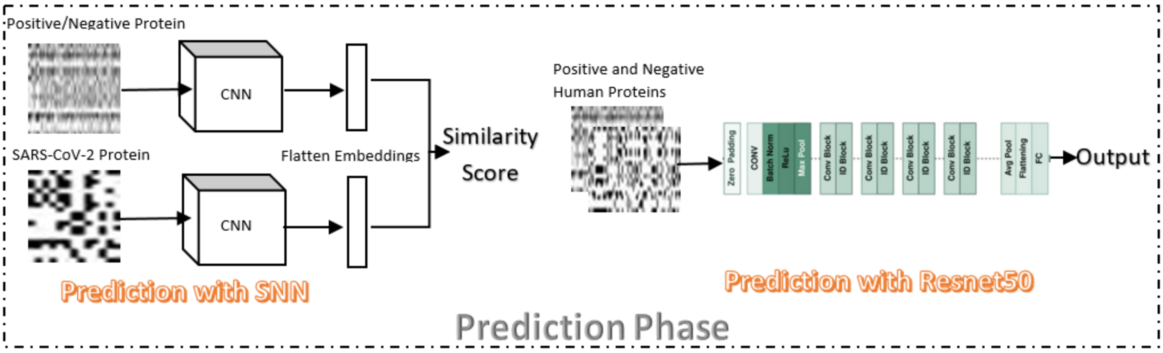

Figure 8. Framework of interaction prediction process.

Table 2. Parameters of networks

Network	#Layers	Batch size	Epochs	Act. func.	Loss func.
SNN	2 2D conv layers, a flatten layer, a dense layer.	16	20	Relu	Contrasive loss
Resnet50	48 conv layers, a max pool layer, an average pool layer.	32	20	Relu	Binary-crossentropy

The prediction phase of the proposed method is given in Figure 8. In the prediction phase, two image classification algorithms were applied and the results were compared. The dataset is divided into train and test sets as 70% and 30% respectively. An SNN network is trained with pairs of images that are related and unrelated to each other. This trained model predicts whether a new pair of images to be given are related to each other. In our problem, "associated" means that a virus and host protein pair are interacting. Therefore, in the training phase, we gave the PSSM images of interacting and non-interacting protein pairs to the SNN. There are 332 interacting protein pairs and 332 non-interacting protein pairs in the dataset. After the network was trained, the SNN learned the common features between these interacting protein pairs. Also, the network learned common features between non-interacting protein pairs. The architecture of Resnet50 is different from SNN. Resnet50 was trained with PSSM images of positive and negative proteins. In the dataset, there are 332 positive proteins and 332 negative proteins. After the network was trained, Resnet50 learned the features of positive and negative proteins.

SNN algorithm was applied as suggested in Chicco [42] using Python environment and Keras library. Resnet50 was implemented using the Matlab environment and Deep Learning Toolbox. Parameters of networks are given in Table 2.

2.7. Evaluation Metrics

To decide whether PSSM matrices are suitable data for detecting interacting proteins, we compared the performances of the algorithms and the performances of PSSM matrices of different sizes according to accuracy, positive predictive value (PPV), negative predictive value (NPV), sensitivity, and F-measure scores. All of these metrics are used to evaluate the classification performance of an algorithm and calculated from the confusion matrix [55]. Confusion matrix diagram is given in Table 3. Accuracy [56] is the ratio of the number of correctly predicted interacting and non-interacting proteins to the total number of samples. Accuracy is a metric that is widely used to measure the success of a model but does not appear to be sufficient on its own. PPV [57], also known as precision, shows how many of the proteins that the model predicts as interacting are actually interacting. NPV [57] shows how many of the proteins that the model predicts as non-interacting are actually non-interacting. Sensitivity [57], also known as recall, measures how many of the proteins known to interact are correctly predicted. F-Measure [58] is calculated by taking the harmonic average of precision and recall values. It is especially important for unevenly distributed datasets. Because it measures not only the error costs of false negatives or false positives but all error costs as well. The mathematical expressions of all these metrics are given in equations 2-6. In the equations, TP, FP, TN, and FN represent true positives, false positives, true negatives, and false negatives, respectively.

Table 3. Confusion matrix

		Prediction Results	
		Positive (PP)	Negative (NP)
Actual Results	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$PPV/Precision = \frac{TP}{TP+FP} \quad (3)$$

$$NPV = \frac{TN}{TN+FN} \quad (4)$$

$$Sensitivity/Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F - Measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

3. Results and Discussion

We investigated whether we could predict the protein-protein interaction network for SARS-CoV-2 from PSSM matrices, which are considered as images within the scope of the study. For this purpose PSSM matrices were converted to grayscale images to train the networks. To decide, which size is appropriate, conversion was done in different sizes. We converted PSSM matrices of positive, negative, and SARS-CoV-2 proteins into 20x20, 50x50, 100x100, 200x200, and 400x400 images, separately. It is clear that big-size images can contain more features but at the same time, big-size images require more computation time.

To compare the performances of binary and multi-class classification approaches, a model for both approaches was applied. The interaction data were labeled as binary classes (positive proteins labeled as 1 and negative proteins labeled as 0) and trained with Resnet50. The output of the network is information whether the relevant human protein interacts with any SARS-CoV-2 protein. SNN behaves like a multi-class classifier by its nature. Because the network receives a pair of images. In our problem, given to the network is PSSM images of SARS-CoV-2 and human proteins. The network output returns information about whether these 2 proteins interacted with each other. This output also acts as a multi-class classifier, as it can tell which SARS-CoV-2 protein interacts with the relevant human protein. In other words, since the inputs given in SNNs determine the classes, there is no need to define multiple classes. In our solution, one of them is a specific protein of a virus and the other one is a host protein. If the output of the network is close to 1, it means that these two proteins are interacting and it also tells which protein of the virus is in interacting.
The data set was randomly divided into train and test sets. The test set ratio was determined as 30%. It was ensured that there were equal numbers of positive and negative samples in the test set. As seen in Table 1, the number of human proteins with which SARS-CoV-2 proteins interact in the data set varies. To avoid the results from being specific to a particular training-test set, we repeated the dividing train-test sets process 5 times and took into account the average results. In order to determine whether the PSSM matrices, which are considered as images, are suitable for detecting interacting protein pairs, image classification was performed with SNN and Resnet50 and their performances were compared. In addition, the performances of different-sized images were also compared to decide on the appropriate image size.

The Table 4 includes a comparison of the average accuracy, PPV, and NPV results of the test phase according to learners and size of PSSM images. Standard deviation values are also given in the Table 4 to examine how the results vary according to different test sets. According to the results, both algorithms can detect virus-host protein interactions with PSSM matrices for SARS-CoV-2. While SNN got the best accuracy results with 200x200 PSSM images, Resnet50 got the best results with 400x400 PSSM images. Although promising results are obtained with PSSM images smaller than 200x200, it can be seen from the results that images of 200x200 and 400x400 sizes contain more distinctive features. PPV is an important metric because it contains information about how many of the proteins that the algorithm finds interactive are actually interactive. NPV on the other hand gives the accuracy rate of proteins labeled as non-interacting. Studies for protein-protein interaction network detection focus on finding interacted proteins rather than non-interacting proteins. One of the advantages of an in silico analysis is to reduce candidate solutions that need to be validated by in vitro or in vivo analyses. For all these reasons, a high PPV value is preferred. SNN achieved better results for big images (200x200 and 400x400) however Resnet50 achieved better results for small images. The standard deviation values calculated according to the results of the test set are generally low. This shows that the algorithms do not obtain very different results compared to the different test sets,

Table 4. Average results for test process

Image Size	SNN			Resnet50		
	Acc	PPV	NPV	Acc	PPV	NPV
20x20	0.805±0.097	0.821±0.122	0.802±0.11	0.844±0.02	0.866±0.044	0.834±0.055
50x50	0.885±0.093	0.867±0.103	0.907±0.08	0.877±0.054	0.874±0.05	0.909±0.112
100x100	0.891±0.054	0.892±0.083	0.896±0.042	0.910±0.024	**0.896±0.026**	0.933±0.069
200x200	**0.915±0.084**	0.893±0.112	**0.957±0.055**	0.840±0.136	0.879±0.111	0.895±0.183
400x400	0.903±0.059	**0.937±0.075**	0.888±0.087	**0.922±0.02**	**0.896±0.024**	**0.960±0.063**

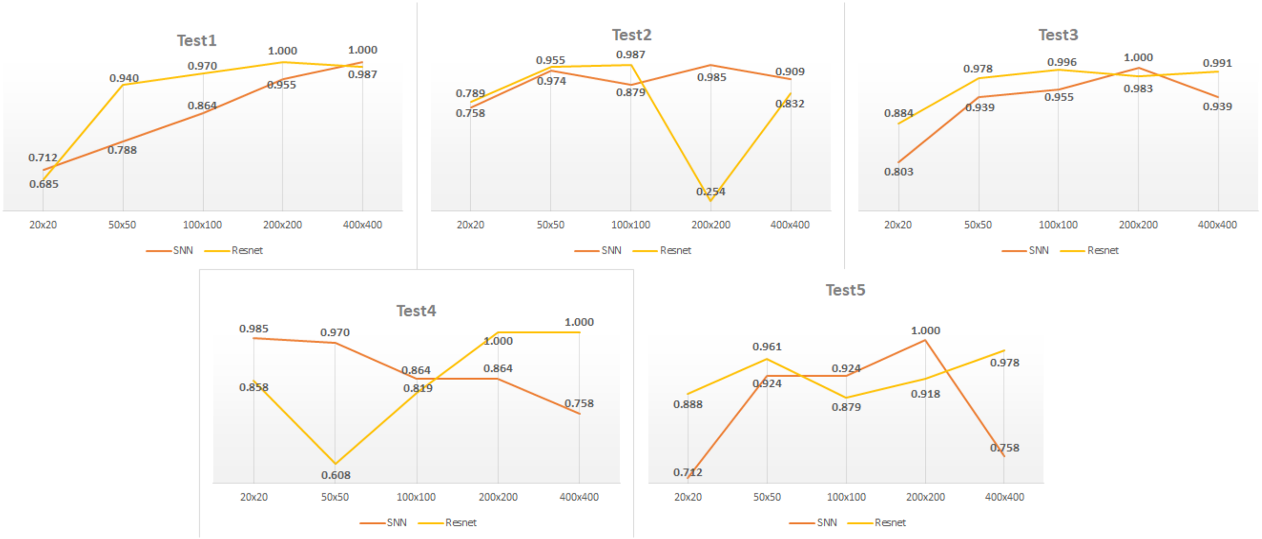

Figure 9. Sensitivity charts according to different test sets

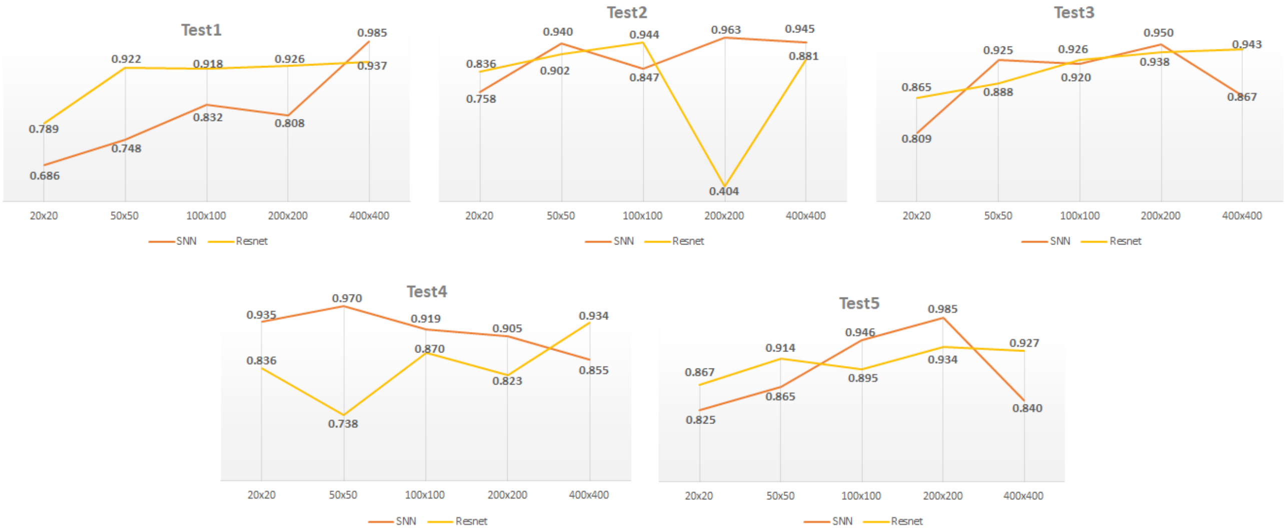

Figure 10. F-Measure charts according to different test sets

that is, the results obtained are not dependent on a specific test set.

The sensitivity and F-measure results obtained from the test sets are given in Figure 9 and Figure 10 respectively. In the charts, the metrics are compared according to the results from each test set. It has been observed that increasing the size does not always increase success for the same PSSM images in a test set. In some test sets like Test 1 and Test 3, the sensitivity value is 1 and the F-Measure value is very close to 1 for 200x200 and 400x400 image sizes. The dataset contains many human proteins that have been confirmed to interact for one SARS-CoV-2 protein, while there are very few human proteins for some SARS-CoV-2 proteins. Because test sets were generated with randomly selected samples, interactions for some SARS-CoV-2 proteins are represented adequately and for others are underrepresented. This situation caused the algorithms not to learn some interaction features well. It is also seen that the algorithms can not obtain results to close each other in the same test set and the same PSSM image size. This is an indication that algorithms learn different features from the same dataset. Resnet's results were more affected by the size of the images, while SNN's results were less dependent on data size.

As mentioned earlier, there are very few in silico studies in the literature to predict protein-protein interaction for SARS-CoV-2. In terms of giving an idea, in Table 5, we compared the results of these studies. Lanchantin et al. [26] used transfer learning, Du et al. [27] used multi-layer perceptron and Dey et al. [31] proposed machine learning algorithms. In the datasets used by Lanchantin et al. [26] and Du et al. [27], there are protein interactions for different viruses as well as SARS-CoV-2. Dey et al. [31], on the other hand, predicted protein interaction network with the dataset used in this study. The machine learning algorithms were trained with sequence-based features extracted from SARS-CoV-2 and human proteins in the dataset. Lee [34] performed a virus type-independent PPI estimation method. They extracted features based on text mining and network similarity from a dataset of interactions of different virus strains with human proteins. They trained the feature dataset with random forest and XGBoost, and performed SARS-CoV-2-human protein interaction prediction with the obtained model. SARS-CoV-2-human interactions were obtained from the IntAct database. The accuracy value of the model developed by the authors for SARS-CoV-2 data was given as 57%. Bell et al. [35] predicted the SARS-CoV-2-human protein interaction network with their pipeline named PEPPI. Estimates were made for 128 interactions obtained from the PSICQUIC27 database. These interactions are between SARS-CoV and/or SARS-CoV-2 and human proteins. The proposed method correctly predicted 94 out of 128 interactions. It can be seen from the results that more successful predictions were obtained with the proposed method.

Table 5. Comparison with recent studies

	F-Measure	Accuracy
Lanchantin et al. [26]	0.114	-
Du et al. [27]	0.867	-
Dey et al. [31]	-	0.723
Lee [34]	-	0.571
Bell et al. [35]		0.734
Proposed Method	**0.880**	**0.922**

The main motivation of this study was the question of whether a model for protein interaction network prediction can be developed with PSSM images. The results obtained show that PSSM images are suitable data for this purpose. The strength of the SNN network is that it can tell which protein of a virus a host protein is interacting with. In other words, the network learns the common features of the virus protein and the host protein with which it interacts. Its high success in predicting positive interactions is another strength of the network. The weakness of the network was its inability to show high performance in learning both negative and positive interactions. As can be seen from Table 4, while it learned only negative interactions well for 200x200 images, it learned positive interactions better for 400x400 images

In cases where experimental data are scarce, a binary classifier developed with a strong network such as Resnet50 can identify whether a host protein interacts with a virus. The weakness of the model developed with Resnet50 is that it cannot tell which virus protein the interaction is with. If the training data can be increased, this deficiency can be eliminated by handling the problem with a multi-class classification approach. However, it will take time to increase the data as the training data are obtained with in-vitro and in-vivo analyzes.

4. Conclusion and Suggestions

The emergence of SARS-CoV-2 virus after SARs-CoV and MERS viruses and causing a pandemic reveals the importance of developing rapid treatment for viral diseases, but it is an indication that similar epidemics may be repeated in the future. In order to understand the behavior of pathogenic viruses that need a host cell to copy their genome, it is necessary to know which protein in the host cell interacts with it. Protein-protein interactions are the main way proteins perform their functions. Therefore, analyzing the protein-protein interactions between the host and pathogenic viruses is useful for understanding the mechanisms of viral infections. In this way, effective antiviral drugs against drug resistance can be designed. Laboratory experiments are needed to definitively identify interacting protein pairs. However, in silico studies can shed light on in vivo analyzes.

Here, an in silico-based prediction of SARS-CoV-2-human protein interaction was performed. Most in silico approaches benefit learning algorithms, and appropriate representations of protein pairs must be obtained in order to make a prediction with learning algorithms. We proposed using PSSM matrices of proteins. The PSSM matrices were converted to grayscale images and SNN and Resnet algorithms were trained with these images. The dataset used for training consists of experimentally confirmed SARS-CoV-2-human protein pairs with which they interact. Because the size of PSSM matrices depends on the length of the protein, different proteins have different sizes of PSSM matrices. A standard scaling was used during the conversion of PSSM matrices to images to obtain images of the same size for all proteins. We converted PSSM matrices into 20x20, 50x50, 100x100, 200x200 and 400x400 size images to decide what size is sufficient to perform a successful protein-protein interaction process. The networks were individually trained with groups of this image size. The randomly selected 30% of the data set was used as the test set. The dataset was divided into train-test sets 5 times with the same procedure. For each training set, models were trained separately and tested with corresponding test set. The average results of the test performances of the models were taken into account in the comparisons. SNN achieved the best average performance with 200x200 PSSM images with an accuracy of 0.915, while Resnet50 achieved the best average performance with 400x400 PSSM images with an accuracy of 0.922. These results showed that protein-protein interaction network prediction can be performed successfully with images obtained from PSSM matrices. The 2 identical subnets contained in the SNN can successfully learn the common properties of interacting and non-interacting human-virus protein pairs. Since PSSM matrices include biological information and some amino acid features, they are used to obtain this information about protein pairs in protein-protein interaction problems. We converted PSSM matrices to images. Obtained results demonstrate that PSSM image approach is useful for predicting interacting protein pairs. It is hoped that the proposed method will provide reference for in vivo studies by applying to protein pairs with unknown interaction status.

Acknowledgment

This work is supported by Turkish Scientific and Technical Research Council-TÜBİTAK with the project number 122E114.

Contributions of the authors

Zeynep Banu Özger: Conceptualization, Writing – original draft, Methodology, Validation, Software, Visualization, Writing – review & editing. Zeynep Çakabay: Validation, Software, Visualization.

Conflict of Interest Statement

The authors declare no conflict of interest

Statement of Research and Publication Ethics

The study is complied with research and publication ethics

References

[1] P. Koehl, “Protein structure similarities”. *Current opinion in structural biology*, 11(3), 348-353, 2001. doi: 10.1016/S0959-440X(00)00214-1.

[2] D. P. Ryan, and J. M. Matthews, “Protein–protein interactions in human disease”. *Current Opinion in Structural Biology*, 15(4), 441-446, 2005. doi: 10.1016/j.sbi.2005.06.001

[3] V. Altuntaş, and M. Gök, “Protein–protein etkileşimi tespit yöntemleri, veri tabanları ve veri güvenilirliği”. *Avrupa Bilim ve Teknoloji Dergisi*, (19), 722-733, 2020. doi: doi.org/10.31590/ejosat.724390.

[4] J. Piehler, "New methodologies for measuring protein interactions in vivo and in vitro". *Current Opinion in Structural Biology*, 15(1), 4-14, 2005. doi: 10.1016/j.sbi.2005.01.008.

[5] S. Xing, N. Wallmeroth, K. W. Berendzen, and C. Grefen, "Techniques for the analysis of protein-protein interactions in vivo". *Plant Physiology*, 171(2), 727-758,2016. doi: 10.1104/pp.16.00470.

[6] S. Vivona, J. L. Gardy, S. Ramachandran, F. S. Brinkman, G. P. S. Raghava, D. R. Flower, and F. Filippini, "Computer-aided biotechnology: from immuno-informatics to reverse vaccinology". *Trends in Biotechnology*, 26(4), 190-200, 2008. doi: 10.1016/j.tibtech.2007.12.006.

[7] S. J. Y. Macalino, S. Basith, N. A. B. Clavio, H. Chang, S. Kang, and S. Choi, "Evolution of in silico strategies for protein-protein interaction drug discovery". *Molecules*, 23(8), 1963, 2018. doi: 10.3390/molecules23081963.

[8] P. Kangueane, and C. Nilofer. "Principles of Protein-Protein Interaction," in *Protein-Protein and Domain-Domain Interactions*. Springer, Singapore, 2018. doi:10.1007/978-981-10-7347-2_8.

[9] B. Parlak, and A. K. Uysal. "On classification of abstracts obtained from medical journals". *Journal of Information Science*, 2020, 46(5), 648-663.

[10] L. Dey, and A. Mukhopadhyay, "A classification-based approach to prediction of dengue virus and human protein-protein interactions using amino acid composition and conjoint triad features," in *2019 IEEE Region 10 Symposium (TENSYMP)*, 2019, June, pp. 373-378, IEEE.

[11] Y. Ma, T. He, Y. Tan, and X. Jiang, "Seq-bel: Sequence-based ensemble learning for predicting virus-human protein-protein interaction". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(3), 1322-1333,2020. doi: 10.1109/TCBB.2020.3008157.

[12] X. Yang, S. Yang, Q. Li, S. Wuchty, and Z. Zhang, "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method". *Computational and Structural Biotechnology Journal*, Vol.18, pp. 153-161, 2020. doi: 10.1016/j.csbj.2019.12.005

[13] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions". *PLoS One*, 7(4), e32289, 2012. doi: 10.1371/journal.pone.0032289.

[14] S. K. Ng, Z. Zhang, and S. H. Tan, "Integrative approach for computationally inferring protein domain interactions". *Bioinformatics*, 19(8), 923-929, 2003. doi: 10.1093/bioinformatics/btg118.

[15] N. Zhang, M. Jiang, T. Huang, and Y. D. Cai, "Identification of Influenza A/H7N9 virus infection-related human genes based on shortest paths in a virus-human protein interaction network". *BioMed Research International*, 2014, 2014. doi: 10.1155/2014/239462.

[16] S. Bandyopadhyay, and K. Mallick, "A new feature vector based on gene ontology terms for protein-protein interaction prediction". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(4), 762-770, 2016. doi: 10.1109/TCBB.2016.2555304.

[17] H. Ge, Z. Liu, G. M. Church, and M. Vidal, "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae". *Nature Genetics*, 29(4), 482-486, 2001. doi: doi.org/10.1038/ng776.

[18] A. Zhang, L. He, and Y. Wang, "Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions". *BMC Bioinformatics*, 18(1), 1-13, 2017. doi: 10.1186/s12859-017-1500-8.

[19] M. D. Dyer, T. M. Murali, and B. W. Sobral, "Computational prediction of host-pathogen protein–protein interactions". *Bioinformatics*, 23(13), i159-i166, 2007. doi: 10.1016/j.patter.2021.100242.

[20] A. Birlutiu, F. d'Alché-Buc, and T. Heskes, "A bayesian framework for combining protein and network topology information for predicting protein-protein interactions". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(3), 538-550, 2014. doi: 10.1109/TCBB.2014.2359441.

[21] S. Erten, X. Li, G. Bebek, J. Li, and M. Koyutürk, "Phylogenetic analysis of modularity in protein interaction networks". *BMC Bioinformatics*, 10(1), 1-14, 2009. doi: 10.1186/1471-2105-10-333.

[22] N. Papanikolaou, G. A. Pavlopoulos, T. Theodosiou, and I. Iliopoulos, "Protein–protein interaction predictions using text mining methods". *Methods*, 74, 47-53, 2015. doi: 10.1016/j.ymeth.2014.10.026.

[23] B. Khorsand, A. Savadi, J. Zahiri, and M. Naghibzadeh, "Alpha influenza virus infiltration prediction using virus-human protein-protein interaction network". *Mathematical Biosciences and Engineering*, 17(4), 3109-3129, 2020. doi: 10.3934/mbe.2020176.

[24] P. Zhou, X. L. Yang, X. G. Wang, B. Hu, L. Zhang, W. Zhang, ... and Z. L. Shi, "A pneumonia outbreak associated with a new coronavirus of probable bat origin". *Nature*, 579(7798), 270-273, 2020. doi: 10.1038/s41586-020-2012-7

[25] A. A. Khan, and Z. Khan, "Comparative host–pathogen protein–protein interaction analysis of recent coronavirus outbreaks and important host targets identification". *Briefings in Bioinformatics*, 22(2), 1206-1214, 2021. doi: 10.1093/bib/bbaa207.

[26] J. Lanchantin, A. Sekhon, C. Miller, and Y. Qi, "Transfer learning with motiftrans-formers for predicting protein-protein interactions between a novel virus and humans". B*ioRxiv*, 36, i659-i667, 2020. doi: 10.1101/2020.12.14.422772.

[27] H. Du, F. Chen, H. Liu, and P. Hong, "Network-based virus-host interaction prediction with application to SARS-CoV-2". *Patterns*, 2(5), 100242, 2021. doi: 10.1016/j.patter.2021.100242.

[28] S. Su, G. Wong, W. Shi, J. Liu, A. C. Lai, J. Zhou, ... and G. F. Gao, "Epidemiology, genetic recombination, and pathogenesis of coronaviruses". *Trends in Microbiology*, 24(6), 490-502, 2016. doi: 10.1016/j.tim.2016.03.003.

[29] B. Khorsand, A. Savadi and M. Naghibzadeh, "SARS-CoV-2-human protein-protein interaction network". *Informatics in Medicine Unlocked*, 20, 100413, 2020. doi: 10.1016/j.imu.2020.100413.

[30] R. Oughtred, C. Stark, B. J. Breitkreutz, J. Rust, L. Boucher, C. Chang, ... and M. Tyers, "The BioGRID interaction database: 2019 update". *Nucleic Acids Research*, 47(D1), D529-D541, 2019. doi: 10.1093/nar/gky1079.

[31] L. Dey, S. Chakraborty and A. Mukhopadhyay, "Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins". *Biomedical Journal*, 43(5), 438-450, 2020. doi: 10.1016/j.bj.2020.08.003.

[32] D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White ... and N. J. Krogan, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing". *Nature*, 583(7816), 459-468, 2020. doi: 10.1038/s41586-020-2286-9.

[33] D. Pirolli, B. Righino, and M. C. De Rosa. "Targeting SARS-CoV-2 Spike Protein/ACE2 Protein-Protein Interactions: a Computational Study". *Molecular Informatics*, 2021, 40(6), 2060080.

[34] H. J. Lee. "An interactome landscape of SARS-CoV-2 virus-human protein-protein interactions by protein sequence-based multi-label classifiers". *BioRxiv*, 2021.

[35] E. W. Bell, J. H. Schwartz, P. L. Freddolino, and Y. Zhang. "PEPPI: Whole-proteome protein-protein interaction prediction through structure and sequence similarity, functional association, and machine learning". *Journal of Molecular Biology*, 2022, 167530.

[36] G. Launay, N. Ceres, and J. Martin. "Non-interacting proteins may resemble interacting proteins: prevalence and implications". *Scientific reports*, 2017, 7(1), 1-12.

[37] R. K. Barman, S. Saha, and S. Das. "Prediction of interactions between viral and host proteins using supervised machine learning methods". *PloS One*, 2014, 9(11), e112034.

[38] T. Sun, B. Zhou, L. Lai and J. Pei. "Sequence-based prediction of protein protein interaction using a deep-learning algorithm". *BMC Bioinformatics*, 2017, 18(1), 1-8.

[39] S.R. Eddy. "Where did the BLOSUM62 alignment score matrix come from?" *Nature Biotechnology*, 2004, 22(8), 1035-1036.

[40] UniProt Consortium. "UniProt: a hub for protein information". *Nucleic Acids Research*, 2015, 43(D1), D204-D212.

[41] J. D. Bernal. "Structure of proteins". *Nature*, 1939, 143(3625), 663-667.

[42] J. C. Jeong, X. Lin, and X. W. Chen. "On position-specific scoring matrix for protein function prediction". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2010, 8(2), 308-315.

[43] R.C. Edgar, and S. Batzoglou. "Multiple sequence alignment". *Current Opinion in Structural Biology*, 2006, 16(3), 368-373.

[44] A. Mohammadi, J. Zahiri, S. Mohammadi, M. Khodarahmi, and S. S. Arab, "PSSMCOOL: a comprehensive R package for generating evolutionary-based descriptors of protein sequences from PSSM profiles". *Biology Methods and Protocols*, 7(1), bpac008, 2022. doi: 10.1093/biomethods/bpac008

[45] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu, "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences". *Bioinformatics*, 31(11), 1857-1859, 2015.

[46] S. Albawi, T. A. Mohammed and S. Al-Zawi. "Understanding of a convolutional neural network", in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, IEEE.

[47] J. Wu, “Introduction to convolutional neural networks”. *National Key Lab for Novel Software Technology*. Nanjing University. China, 5(23), 495, 2017.

[48] S. Balaji, S. (2020, Aug 29). “Binary Image classifier CNN using TensorFlow”, *medium.com*. Aug. 29, 2020. [Online]. Available: https://medium.com/techiepedia/binary-image-classifier-cnn-using-tensorflow-a3f5d6746697. [Accessed: 15/11/2022].

[49] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition”, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, IEEE, pp. 770-778.

[50] P. Roy, S. Ghosh, S. Bhattacharya and U. Pal. “Effects of degradations on deep neural network architectures”. A*rXiv,* abs/1807.10108, 2018

[51] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database”, in *2009 IEEE Conference on Computer Vision and Pattern Recognition,* June, 2009, pp. 248-255, IEEE.

[52] D. Chicco, “Siamese Neural Networks: An Overview”, in: *Cartwright, H. (eds) Artificial Neural Networks. Methods in Molecular Biology*, vol 2190. Humana, New York, NY, 2021. doi:10.1007/978-1-0716-0826-5_3.

[53] L. Hudec, and W. Bencsova, “Texture similarity evaluation via siamese convolutional neural network”, in *2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP),* June, 2018, pp. 1-5, IEEE.

[54] M. D. Li, K. Chang, B. Bearce, C. Y. Chang, A. J. Huang, J. P. Campbell, ... and J. Kalpathy-Cramer. “Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging”. *NPJ Digital Medicine*, 2020, 3(1), 1-9.

[55] J. Liang. “Confusion matrix”. *POGIL Activity Clearinghouse*, 2022, 3(4).

[56] S. V. Stehman. “Selecting and interpreting measures of thematic classification accuracy”. *Remote sensing of Environment*, 1997, 62(1), 77-89.

[57] H. B. Wong and G. H. Lim. “Measures of diagnostic accuracy: sensitivity, specificity, PPV and NPV”, in *Proceedings of Singapore Healthcare*, 2011, 20(4), 316-318.

[58] D. M. Powers. “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. *ArXiv preprint arXiv:2010.16061*, 2020.