Journal of Digital Science

ISSN 2686-8296

Volume 3 Issue 2 **December 2021**

CONTENTS

Coronavirus Genome Sequence Similarity and Protein Sequence Classification

Partha Mukherjee[1[0000-0001-9244-8600]],
Youakim Badr[1[000-0002-8976-7894]],
Srusthi N. Karvekar[1[0000-0001-8913-3368]],
Shanmugapriya Viswanathan[1[0000-0003-4459-2274]]

[1]The Pennsylvania State University, Great Valley, Malvern, PA-19335, USA

https://doi.org/10.33847/2686-8296.3.2_1

Received 26.10.2021/Revised 01.11.2021/Accepted 09.12.2021/Published 28.12.2021

Abstract. The world currently is going through a serious pandemic due to the coronavirus disease (COVID-19). In this study, we investigate the gene structure similarity of coronavirus genomes isolated from COVID-19 patients, Severe Acute Respiratory Syndrome (SARS) patients and bats genes. We also explore the extent of similarity between their genome structures to find if the new coronavirus is similar to either of the other genome structures. Our experimental results show that there is 82.42% similarity between the CoV-2 genome structure and the bat genome structure. Moreover, we have used a bidirectional Gated Recurrent Unit (GRU) model as the deep learning technique and an improved variant of Recurrent Neural networks (i.e., Bidirectional Long Short Term Memory model) to classify the protein families of these genomes to isolate the prominent protein family accession. The accuracy of Gated Recurrent Unit (GRU) is 98% for labeled protein sequences against the protein families. By comparing the performance of the Gated Recurrent Unit (GRU) model with the Bidirectional Long Short Term Memory (Bi-LSTM) model results, we found that the GRU model is 1.6% more accurate than the Bi-LSTM model for our multiclass protein classification problem. Our experimental results would be further support medical research purposes in targeting the protein family similarity to better understand the coronavirus genomic structure.

Keywords: Coronavirus Disease of 2019 (COVID-19), Severe Acute Respiratory Syndrome (SARS), Genome Structure, Basic Local Alignment Search Tool (BLAST), Gated Recurrent Unit (GRU), Protein Family Accession.

1. Introduction

1.1. Background

In December 2019, an acute respiratory disease caused by a newly identified beta(β)-coronavirus, occurred in Wuhan, China and received attention all over the world. Initially, this coronavirus was named as the 2019-novel coronavirus (2019-nCoV) on January 12, 2020 by the World Health Organization (WHO). Coronavirus Study Group (CSG) of the international committee proposed to name the new coronavirus as SARS-CoV-2 whereas WHO officially named the disease as coronavirus disease 2019 (COVID-19) on February 11, 2020 [1]. On 7 January 2020, Chinese scientists rapidly isolated a SARS-CoV-2 from a patient to isolate the genome sequence of the SARS-CoV-2 [2].

SARS-CoV is a member of the Coronoviridae family of enveloped, positive stranded RNA viruses, which have a broad host range. Coronavirus infections in rodents, cats, pigs and cattle can be responsible for enteric diseases or cause respiratory diseases in people, cattle and birds. Twenty-three putative proteins,

including four major structural proteins; nucleocapsid (N), spike (S), membrane (M), and small envelope (E) encode from the 27-32 kb genomes of coronavirus. Out of these, the spike protein on the viral surface facilitates the entry and viral attachment to the host cell. In addition, variations of S protein among strains of coronavirus are responsible for host range and tissue tropism. However, the S, M, and N mature proteins all contribute to generate the host immune response as seen in transmissible gastroenteritis coronavirus, infectious bronchitis virus, pig respiratory coronavirus, and mouse hepatitis virus [3].

Among the several coronaviruses that are pathogenic to humans, most are associated with mild clinical symptoms, with two notable exceptions: 1) severe acute respiratory syndrome (SARS) coronavirus (SARS-COV); and 2) Middle East Respiratory Syndrome (MERS) coronavirus (MERS-COV). SARS-COV was a novel beta-coronavirus that emerged in Guangdong, southern China in November, 2002. It resulted in more than 8000 human infections and 774 deaths in 37 countries during 2002–03. MERS-COV was first detected in Saudi Arabia in 2012 and was responsible for 2494 laboratory-confirmed cases of infection and 858 fatalities since September 2012, that included 38 deaths following an outbreak into South Korea [3].

1.2. Origin and Transformation

The SARS-CoV-2 is a β-coronavirus, which is enveloped non-segmented positive-sense RNA virus. Coronaviruses (CoV) are divided into four genres, including alpha, beta, gamma, and delta. Alpha-CoV and beta-CoV can infect mammals, while gamma-CoV and delta-CoV tend to infect birds. Previously, six CoVs have been identified as human-susceptible virus strains, among which alpha-CoVs such as: 1) HCoV-229E, and 2) HCoV-NL63; and beta-CoVs such as: 1) HCoV-HKU1, and 2) HCoV-OC43 with low pathogenicity cause mild respiratory symptoms similar to a common cold respectively. In HCoV-XXXX, the pattern "XXXX" represent the strain names such as 229E, HKU1, OC43 etc. The other known beta-CoVs, SARS-CoV and MERS-CoV lead to severe and potentially fatal respiratory tract infections [4]. The genome related terminologies are listed in the glossary shown in Table 1 below.

In this study, we perform genome structure analysis and compare the level of similarity among genome structures of coronavirus genomes isolated from a COVID-19 patient, SARS patient and a bat.

Further we classify the protein structure of all the genome structure into different protein family accession to identify prominent protein family accession present in all the three viruses. The protein accession number is an identifier for family of proteins with similar functions and structure. Each family has an accession number named by pfam / uniport.

The remaining sections of this study is arranged as follows: In section 2, we investigate the related works performed as part of medical and technological studies in alignment with the COVID-19 virus. In section 3, we present and conduct the genome structure analysis in order to find prominent protein families in the viruses. We preprocess the dataset and perform data exploratory analysis in section 4. In section 5, we discuss the experimental results and their significance with respect to answering our research questions. Finally, we conclude our work and identify future work in section 6.

Table 1. Glossary of terms

GLOSSARY
Potive-Stranded RNA viruses Some viruses, such as coronaviruses, carry their genetic material as RNA rather than the more typical DNA-based genomes. Positive stranded RNA (also called plus-stranded) indicates that the single stranded RNA genome is of the same sense as coding messenger RNA
Beta (β)-Coronavirus It is in the subfamily Orthocoronavirinae in the family Coronaviridae, of the order Nidovirales. They are enveloped, positive-sense, single-stranded RNA viruses of zoonotic origin.
SARS-CoV The strain of coronavirus causing global outbreak of contagious and sometimes fatal respiratory illness which appeared in China in 2002.
MERS-CoV The coronavirus causing respiratory illness which appeared in the Arabian Peninsula in 2012.
CODON A sequence of 3 bases in a m-RNA strand which corresponds to a amino acid. (Eg: UAA, AUG, etc.)
Open Reading Frames (ORF) A continuous stretch of codons that begins with a start codon (usually AUG) and ends at a stop codon (usually UAA, UAG or UGA)
Non-Structural Proteins (NSP) A protein encoded by a virus but that is not part of the viral particle.
HCoV Refers to Human coronavirus
MetaTranscriptomic Refers to the science that studies gene expression of microbes within natural environments

2. Literature Review

Coronaviruses either leads to respiratory diseases or enteric infections in various animal species. These viruses can also cause hepatic and neurological diseases. Human coronaviruses, identified in the 1960s, are responsible for up to 30% of respiratory infections with prototypes HCoV-OC43 and HCoV-229E. The three serotypes of coronavirus have three main classes identified by the phylogenetic analysis [5].

A large number of studies have proved that the pathogen of COVID-19 is a novel coronavirus till date. The COVID-19 pathogen has a linear single-stranded positive-strand RNA genome of about 30 kb that belong to the Beta (β) Coronavirus genus and Sarbecovirus subgenus [6].

The complete genome isolated from a COVID-19 patient who worked in the Wuhan sea-food market, has one strain of size 29.9 kb [5]. SARS-CoV and MERS-CoV have positive-sense RNA genomes of 27.9 kb and 30.1 kb, respectively [7]. It has been stated that the genome of coronavirus contains a variable number, up to six to eleven, of Open Reading Frame (ORFS). Most of the RNAs located in the first ORF translates two polyproteins and encodes 16 Non-Structural Proteins (NSP). This forms two-thirds of the viral RNA. The remaining ORFs encode accessory proteins, that

interfere with the host's innate immune response and structural proteins. The four essential structural proteins include spike (S) glycoprotein, small envelope (E) protein, matrix (M) protein, and nucleocapsid (N) protein [2].

Deep Meta-Transcriptomic sequencing on genome isolated from Wuhan patient, performed by Wu et al. [8] contained 16 predicted NSPs. When compared with the SARS-Cov and MERS-Cov, this genome sequence was closer to the SARS-like bat coronavirus. At the protein level, there are no amino acid substitutions that occurred in NSP7, NSP13, with the exception in NSP2, NSP3, spike protein. NSP(XX) are the names isoforms of the NSPs, (XX is the integer number) where the naming is done chronologically and the numbers (i.e., XX) may not have any significance on protein structures [8]. Another recent research suggested that the mutation in NSP2 and NSP3 play a role in capability of infection, and differentiation mechanism of SARS-CoV-2 [9]. These studies influence the researchers for exploring the difference in the host and transmission between the SARS-CoV-2 and SARS-CoV and consequently opens avenue of further exploration to find the potential therapeutic targets [5].

Ruan et al. [3] analyzed the genotypes of COVID-19 in different patients from several provinces and found that SARS-CoV-2 had been mutated in different patients in China. Tang et al. [10] conducted a population genetic analyses of 103 SARS-CoV-2 genomes and classified out two prevalent evolvement types of SARS-CoV-2, L type (~ 70%) and S type (~ 30%). The L type strains are derived from S type and are more contagious in nature. Thus, this novel coronavirus needs to be monitored closely to inspect this pandemic.

Full-genome sequence analysis of COVID-19 revealed that SARS-CoV-2 belongs to beta (β) coronavirus, but it is divergent from SARS-CoV and MERS-CoV that caused epidemics in the past. The COVID-19 along with the Bat-SARS-like coronavirus forms a distinct lineage within the subgenus of the Sarbecovirus [11].

One of the long-term problems in molecular biology is to understand the relationship between an amino acid sequence and the protein function. This insight will result in useful scientific implications. As discussed in Bileschi et al [12], the classification of 17000 protein families is not feasible with conventional machine learning techniques like SVM, Decision Trees and ensemble models. Conventional machine learning models have limitations such as the requirement of substitution matrices, hard tuned scoring functions, feature engineering and sequence alignment to mention a few. By leveraging advanced Artificial Intelligence (AI) technology and especially Deep Learning methods, we overcome these constraints and directly predict the protein functional annotations from the data. Deep Learning algorithms efficiently co-locates sequences from unseen families with more accuracy [13]. The Deep Learning model learns the relationship between unaligned amino acid sequences and their functional annotations across protein families of the Pfam database [14]. Deep sequencing technology facilitates parallel processing of numerous distinct genome fragments and identify millions of base pairs within a few hours. The genomic analysis using deep sequencing determines: a) the structure and location of genes, b) regulatory elements, c) non-coding RNAs, and predict the gene functions [15]. For these reasons, we propose Deep Learning based model GRU to predict the protein family accession number with the protein sequence and compare it with the performance of Bi-LSTM method [16-18] using the same dataset. Our model captures the patterns between the sequence formation and protein function that cannot be easily identified. To the best of our knowledge, the GRU technique in this study has not been explored and thus can be used for creating the protein profile of the newly emerged coronavirus.

3. Research in Context

In the previous section, we described the genomic characteristics of COVID-19 with similarities and differences to other coronaviruses, including the virus that caused the SARS epidemic in 2002–03. We also classified the proteins found in the virus into various families to help identify its functions. Discovering the functions of new proteins not only allows one to better understand their roles in their native contexts, but also utilize them in synthetic biology to assemble new biological circuits and pathways for useful applications such as production of valuable drugs for treating the disease.

In this study we explore the following research questions:

1. What are the characteristics of the genome sequence of COVID-19 coronavirus?

2. What are the similarities and differences between COVID-19 coronavirus, the bat-coronavirus, and SARS-coronavirus genome structures?

3. Is the COVID-19 coronavirus more similar to bat coronavirus or is it mutated SARS coronavirus?

4. Which class of the amino acid of the protein family accession does protein belong to?

For our analysis we use the Basic Local Alignment Search Tool (BLAST) algorithm to find similarity with other genome sequences. BLAST is one of the most heavily used sequence analysis tools available in the public domain [19, 20]. There is now a wide choice of BLAST algorithms that can be used to search many different sequence databases via the BLAST web page (http://www.ncbi.nlm.nih.gov/BLAST). The algorithm–database combinations can thus be executed either with default parameters or with customized settings. The results can also be viewed in various ways [21] and includes various calculations such as the highest score of the sequence, the total score across the genome, query coverage, expected value and percentage identity.

Since our aim is to find if the novel coronavirus has originated from bat or is mutated SARS coronavirus, we use deep learning methods to identify the protein accession families. Based on the outcome, we are able to understand the initial protein profile for the complete virus and our analysis could be useful to initiate drugs discovery research in newly identified organisms.

4. Experimental Set-up

4.1. Datasets

SARS-CoV-2 virus is a beta coronavirus, like MERS-CoV and SARS-CoV, both of which have their origins in bats. The Chinese rufous horseshoe bat *(Rhinolophus sinicus)* has been suggested to have direct lineage to the SARS coronavirus (SCoV), and the diversity of SARS-like CoVs (SLCoV) [22]. Hence the genomic and protein profiles of a COVID-19 and SARS causing coronavirus is compared to that isolated from a bat. The GenBank/EMBL/DDBJ accession numbers for the sequences obtained in our study are LR757996, DQ182595, and FJ588686. The genomic data used are coronavirus sequence assembly derived from patients affected in Wuhan in 2019-2020, patients infected with severe acute respiratory syndrome (SARS) in Zhejiang in 2003 and *Rhinolophus sinicus* species in 2006. The *SARS-CoV-2 strain (Genbank:* LR757996) is the coronavirus sequence assembly isolated from COVID-19 patient in 2019-2020 Wuhan outbreak. The SARS-COV strain (Genbank: DQ182595) is isolated from the throat swab from the first patient with SARS in Zhejiang. The bat SARS-CoV strain (Genbank: FJ588686) is isolated from *Rhinolophus sinicus* and is phylogenetically closer to human SARS-CoV than virus strains from other bat species

[18]. Viral sequences were downloaded from NCBI nucleotide sequence database [23].

Data Preparation

Protein profiling was conducted based on the pfam data of protein sequences and their family accession numbers. The dataset consists of 1.33 million protein sequences belonging to 17929 protein families collected from Kaggle. The data included sequences of length up to 2000 amino acids as depicted in the left sub-figure of Fig. 1.

We explore the sequence length of the protein sequences and plot the frequency of amino acids in Fig. 1. The histogram of the sequence length exhibits right skewness on count of protein sequences with the average sequence length of 155 amino acid and the median of 119 amino acid. For the purposes of this research, the sequences were limited to length of 119 amino acids to remove outliers. Also from the frequency of the amino acids in the dataset, we observe that the leucine (L) has the highest frequency in the whole dataset followed by Alanine (A) and Valine (V).

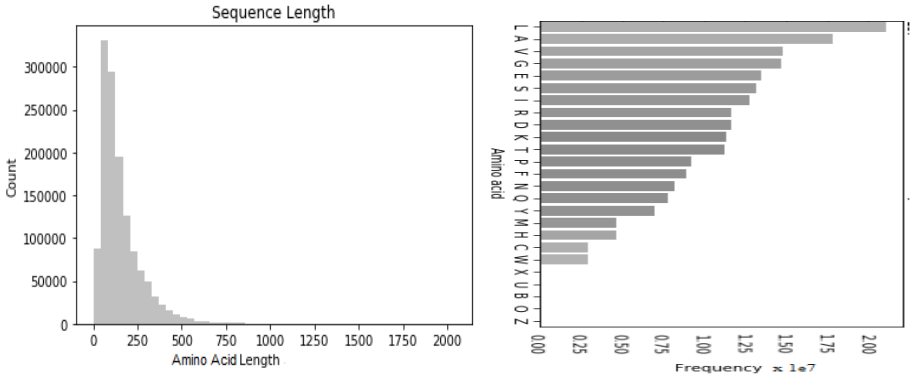

Fig. 1. Protein sequence length and frequency of Amino acids. Source: ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/Pfam-A.seed.gz, https://www.kaggle.com/googleai/pfam-seed-random-split

The dataset was split into training, validation and testing using the 80:10:10 ratio. For computational purposes, the number of families trained was limited to 1,000 common families which approximately correspond to 40% of the dataset. The protein sequences are integer coded for the exhaustive set of all the 20 amino acids [16].

4.2 Genomic Profiling

The Genomic Sequences are combinations of four nucleotides named: A(Adenine), T(Thymine), G(Guanine) and C(Cytosine). The combination of these sequences form the three base pair code for 20 different known amino acids. The amino acids sequences form the linear protein structure.

Subsequently, the genomes are compared for sequence similarity using the Basic Local Alignment Search Tool (BLAST). As the algorithm is based on finding local alignment, there may be multiple discrete regions of sequences with higher similarity score. BLAST is a sequence similarity search program that can be used via a web interface or as a stand-alone tool. There are several types of BLAST algorithms to compare all combinations of nucleotide or protein queries with nucleotide or protein databases. BLAST relies on heuristics to find short matches between two sequences and attempt to start alignments from these 'hot spots.' In addition to performing

alignments, BLAST provides statistical information to decipher the biological significance of the alignment; this is the 'expected' value, or false positive rate.

The Graphic Summary shows alignments of our query sequences. Fig. 2 represents how the similarity score is calculated. The aligned or positively matched sequence have a corresponding value and the gap represents a penalty. The difference between the summation of the values and summation of the penalties is used to compute the sequence similarity between genomes. The similarity score represents maximum sequence similarity.

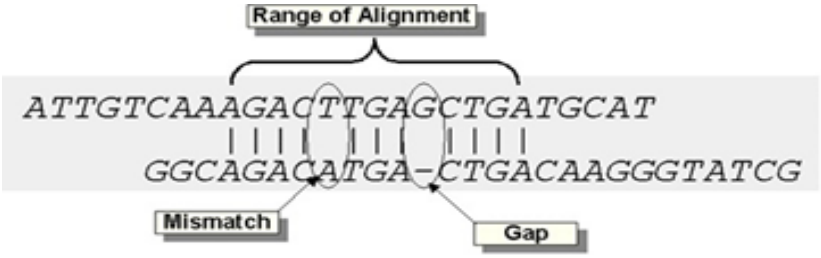

Fig. 2. Similarity Score

S is the sequence similarity score and is defined as follows:

$$S = \sum(identities, mismatches) - \sum(gap\ penalities) \quad (1)$$

$$Score = Max(S) \quad (2)$$

$$\%\ Coverage = \frac{pairwise\ alignment\ score\ of\ subject\ sequence(Score)}{length\ of\ query\ sequence} * 100 \quad (3)$$

4.3. Methodology

4.3.1. Gated Recurrent Unit (GRU)

Artificial Neural Networks (ANN) have been widely used in the field of text mining [24]. We have used Gated Recurrent Unit (GRU) [25] as the deep learning methodology in our study. GRU is a variation of Long Short-Term Memory (LSTM) and an improved variant of Recurrent Neural Networks (RNN) [26]. GRU has achieved remarkable performance in text classification problems [27]. Protein family is predicted from the unaligned protein domain sequence using a GRU due to its high success rate in identifying predominant protein family from the sequences [28, 29]. Our study can identify potential targets for combatting viral infection.

Neural Networks which is a subfield of artificial intelligence (AI) seek to build predictive models to classify or get insight from different types of data. Neural Network consists of different computational layers, including input, output layer, and hidden layers. The hidden layers are connected to input and output neuron on either side. An activation function is applied to generate output for an input. The architecture of Convolutional Neural Network (CNN) is different as it has a set of filters which scan the input for features irrespective of their position in the sequence. In Recurrent Neural Network, there is a time-delayed connection in between the neurons along with the feed forward neural network. For a biological sequence, RNN considers one sequence at a time and transfers information from output of previous step's hidden layer to input layer of next step. Moreover, in RNNs, the gradient of the loss function

exponentially decays over time (vanishing gradient). The GRU models are a varied type of RNN which is used for problems involving protein sequences [30, 31]. GRU has forget gates like LSTM architecture but with fewer parameters as GRU does not have output gate. The memory cells in LSTM allow it to learn longer-term dependencies. Along with it, it filters out the memory to store important input for many steps. There is, of course, a higher complexity and larger computer cost involved in the LSTM model. A GRU would be superior because it is simpler yet retains the ability to train for long-term dependency. GRUs are also easier to modify and can be trained faster than LSTMs. In this study we also present the Bi-LSTM result along with the GRU model performance.

In Fig. 3 we provided the internal architecture of GRU and Bi-LSTM models. In Figure 3(a) we show the building block of GRU where each building block can monitor the flow of information by updating and resetting gates revealing the memory at every step. The GRU attains the equilibrium between the previous and the next states of the memory as its output.

When both the input and output are sequences, we use bi directional GRU model which is an extended GRU model. In such cases, recurrence occurs in the input from both forward and backward directions of network. The use of bidirectional inputs can operate in two ways, one from past to future and one from future to past, allowing it to retain contextual knowledge from past and future at any time. In our study, the

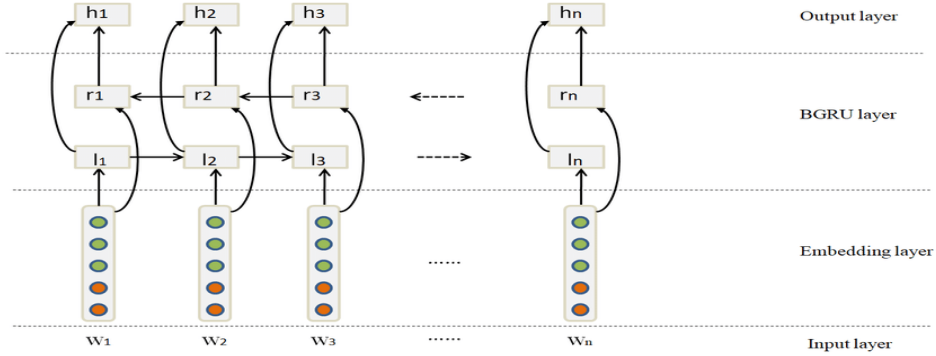

Fig. 3a. Architecture of GRU. Source: based on [32]

input is an amino acid sequence which can be processed from N-terminus to C-terminus and C-terminus to N-terminus. The hidden layer of this model handles the protein sequences by capturing the dependency information of subsequences from all the intermediate hidden values. The feature of the target subsequence in a cell is calculated based on its dependencies between the left neighboring subsequence and the right neighboring subsequence. This way the learning from the next step can be utilized to predict the earlier steps of the model. This leads to develop bidirectional RNN where neurons are divided into forward and backward layers. As sequencing and experimental characterization of data increase rapidly, bidirectional GRU could be useful over bi directional RNN, for discovery and prediction of similarity for a wide range of protein functions. Bidirectional GRU also fixes the problem of vanishing gradients inherent in regular RNNs.

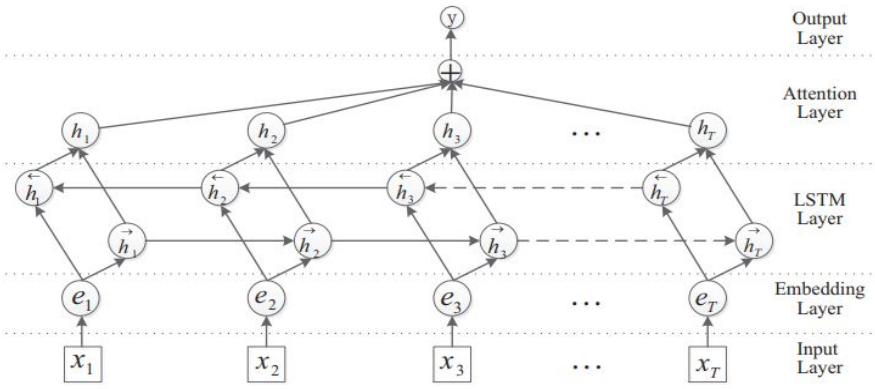

Fig. 3b. Architecture of Bi-LSTM Models
Source: https://github.com/kwonmha/Bidirectional-LSTM-with-attention-for-relation-classification

The mathematical abstraction of hidden layer activation in a neuron unit of GRU are defined as:

$$l_t = \sigma(W_l x_t + U_l h_{t-1} + b_l)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$
$$p_t = tanh(W_h x_t + U_h[r_t * h_{t-1}] + b_r)$$
$$h_t = (1 - l_t)h_{t-1} + lp_t$$

Where $l_t \rightarrow update\ gate$
$r_t \rightarrow reset\ gate$
$p_t \rightarrow candidate\ activation\ vector$
$h_t \rightarrow output\ memory\ state$
$h_{t-1} \rightarrow output\ of\ Previous\ GRU\ block$
$W_x \rightarrow weights\ of\ respective\ gate(x)\ neurons$
$\sigma \rightarrow sigmoid\ function$
$U_x \rightarrow recurrent\ weight\ matrics\ for\ gate(x)\ neurons$

The mathematical abstraction of hidden layer activation in a neuron unit of Bi-LSTM are defined as:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$$
$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$
$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o)$$
$$\hat{c}_t = tanh(W_c.[h_{t-1}], x_t) + b_c$$
$$c_t = f_t * c_{t-1} + i_t * \acute{c}_t$$
$$h_t = o_t * \tanh(c_t)$$

Where
$i_t, f_t\ and\ o_t\ denote\ input\ gate, forget\ gate$ and $output\ gate$
$\sigma \rightarrow sigmoid\ function$
$W_x \rightarrow weights\ of\ respective\ gate(x)\ neurons$
$h_{t-1} \rightarrow output\ of\ previous\ LSTM\ block(at\ t-1)$
$x_t \rightarrow input\ at\ current\ timestamp$
$b_x \rightarrow biases\ for\ the\ respective\ gate(x)$
$c_t \rightarrow cell\ state\ (memory)\ at\ timestamp\ (t)$
$\acute{c}_t \rightarrow candidate\ for\ cell\ state\ (at\ timestamp\ t)$

4.3.2. Model Implementation and Training

The GRU model is trained on 439,493 protein sequences using "Adam" optimizer for 10 epochs with the batch size of 256 sequences. The model is tested and validated with 54,378 sequences. The model architecture comprises of an

embedding layer, a dropout layer at 0.5 rate, a GRU layer with 1024 units and a last dense layer with 1000 neurons to classify into the 1000 families. As a multiclass classification problem, we classify a given amino acid sequence into one of 1000 protein families using the 'softmax' activation function. In order to compare their performances, we have used the same parameters to train the Bi-LSTM model.

5. Results and Discussion

5.1. Genome Similarity Analysis

The genome structure of the three viruses isolated from three different sources namely COVID patient, SARS patient, and the virus isolated from a bat are further partitioned into different base pairs that the DNA structure consists of. This base pairs define different pairing relationships that the DNA sequence generates. We can see from Table 2 constructed from exploratory analysis of the dataset, the COVID structure has a genome length of 29,868, followed by SARS structure with a length of 29,706, and finally by bat structure with a length of 29,059. Thus, all three structures have a substantial amount of genome length for our research to check for the similarity between their structures. Further, we studied the composition of the four nucleotides which make up the genome sequence. The genome wide exploration of the sequence showed little difference between the strains with close G-C content as well as the nucleotide percentages in the following order T>A>C>G. The similar G-C content is accounted for as a factor for the structural similarity between the strains.

Table 2. Table listing the genomic comparison

	Covid	SARS	Bat
Genome length	29,868	29,706	29,059
T(Thymine)	9,586	9,135	8,852
A(adenine)	8,933	8,450	8,266
G(Guanine)	5,861	6,184	6,085
C(Cytosine)	5,488	5,937	5,856
G-C Content	38.00%	40.80%	41.10%

We resorted to global alignment which is a sequence alignment over the entire length of two or more nucleic acid or protein sequences using Needleman-Wunsch algorithm [33]. Fig. 4 shows the pairwise alignment of the sequences based on global alignment algorithm. We found 82.42% identity match between the COVID-19 patient and the virus that has been isolated from a bat using the NCBI BLAST. Based on the NCBI similarity score of 27,375 using equation 1, we compute the coverage score of 91.6% between the genomes by equation 3.

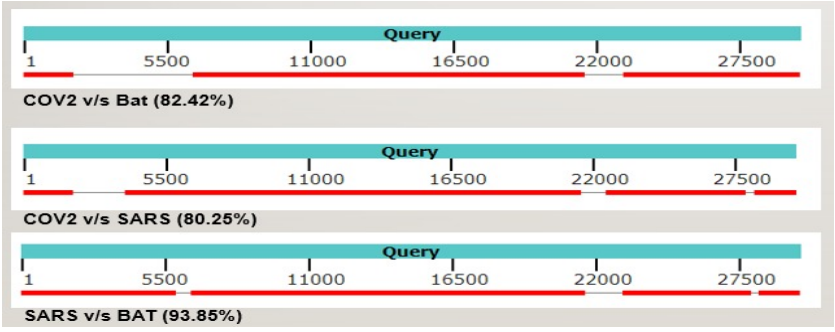

Fig. 4. Pairwise Alignment

In Fig. 4, all the highlighted red areas show different base pairs positions that have matched with most of the genome similarities between the viruses isolated from bats, SARS and the COVID-19 patient. In the top segment of figure 4, the highlighted red areas from 6000bp to 20000bp position on the genome show different base pairs positions that have matched with most of the genome similarities between the viruses that are isolated from bat and the COVID-19. In the middle segment of figure 4, the highlighted red areas from 4000bp to 20000bp show different base pairs positions that have matched with most of the genome similarities between the viruses isolated from SARS and the COVID-19 patients. From the above-received percentage similarities (82.42%), we can infer that CoV-2 has evolved from bat virus. Further the analysis of genome similarity between the viruses that are isolated from bats and the SARS patient as shown in the bottom segment of the figure 4, we observe a staggering similarity of 93.85%. We can see from the highlighted red areas that the entire genome structure is a match except for the areas between positions 6000bp to 6500bp and 21000bp to 23000bp.

5.2. Model Analysis

As depicted in Fig. 5, the proposed bidirectional GRU model is trained over the samples with known protein families from Pfam database and performed well through the epochs with no sudden spikes or drops in either training or validation. Further, there is a sharp significant increase and decrease in the accuracy and loss respectively at 2nd epoch. The accuracy curves, started from 18 % of training data and 91% of validation data in the first epoch. It goes up to 99% for training data and 98% for the validation data at 8th epoch (i.e. before reaching 10th epoch). On the other hand Figure 6 exhibits the curves for Bi-LSTM model, that started from 35% in the first epoch on training data and 69% on validation data, reached to a level of 94% and 96% for training and validation dataset respectively by the end of 10th epoch. Figure 5(b) shows that for GRU model there was significant levels of drop in loss values from 4.1 to 0.01 for training data and 0.34 to 0.06 for the validation data. Figure 6(b) shows the loss reduction for Bi-LSTM model from 3.3 to 0.247 for training data and from 1.6 to 0.129 for validation data respectively.

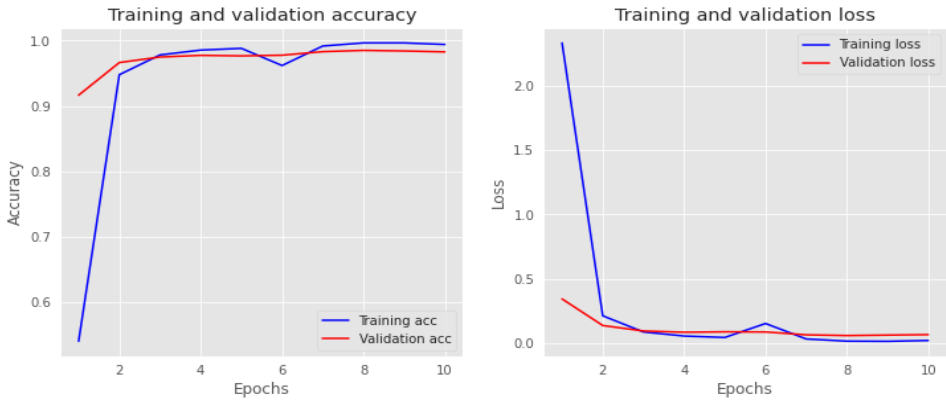

Fig. 5. Accuracy and Loss curves for GRU model

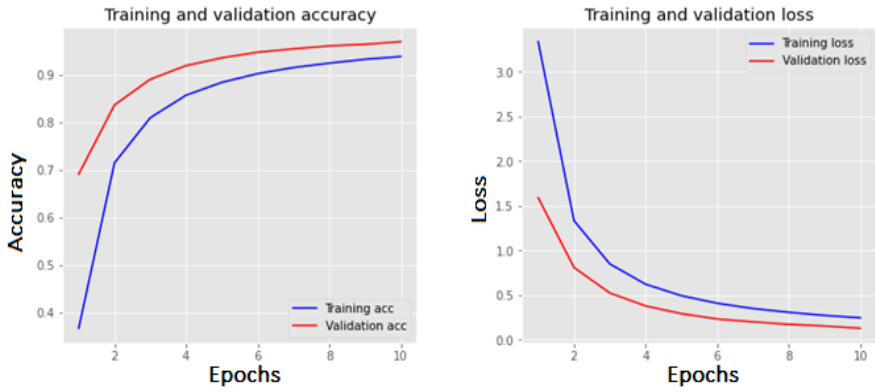

Fig. 6. Accuracy and Loss curves for Bi-LSTM model

Our proposed GRU model achieves training accuracy of 99.6%, validation and test accuracies of 98.2% compared to the Bi-LSTM model with training accuracy of 93.8%, validation and test accuracies of 96.6% as shown in Table 3. There is an improvement of 1.6% of our GRU model over the Bi-LSTM model for both the test and validation dataset and ≈6% improvement over training accuracy. The models use the Pfam database of all the 1000 protein families. This does not include the protein families belongs to SARS-CoV-2 as they are not identified for COVID-19 virus. Protein families are group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure. Thus, the acquired GRU model is capable of finding different protein family accessions present in different genome sequences that are included in Pfam database. Since the GRU model uses the unaligned protein sequences to predict families, by predicting the protein families of CoV proteins, we will be able to identify functionally similar proteins. These functionally similar proteins can give an insights on the pathogenicity of the viruses.

Table 3. GRU and Bi-LSTM Model Performance Metrics

	TRAIN		**TEST**		**VALIDATION**	
Model	GRU	Bi-LSTM	GRU	Bi-LSTM	GRU	Bi-LSTM
Data Size	439493	439493	54378	54378	54378	54378
Accuracy	99.6%	93.8%	98.2%	96.6%	98.2%	96.6%
Loss	0.01	024	0.06	0.13	0.06	0.13

The CoV-2 protein sequences are translated from the DNA sequences downloaded from NCBI. These CoV-2 protein sequences generated are used to predict the protein families. We got an average of 76% probability as the measure of confidence for the prediction of protein families included in CoV-2. We computed the result from the average probabilities from the probability distribution against the protein families generated by the Softmax function used in the trained model. Our model classifies protein family accessions which are respectively present in the genome structure of SARS-COV-2, SARS-COV and bats. Table 4 defines the count of total amino acids, total proteins, number of functional proteins, and the number of family classes present in each classified genome structure.

5.3. Protein Family Accession Analysis

Table 4 exhibits the profile of the protein sequences of three different strains. Further stacking of protein family accession based on the total number of occurrences

gives a clear distinction of which protein family accession is a prominent member of all three viruses as shown in Figure 7. We found that the PF00560.3 family is the prominent protein family accession in all three viruses. PF00560.3 is a family of leucine-rich repeats and a similar motif has been discovered to play an important role in pathogenesis of SARS-CoV [34]. The presence of similar proteins indicates nucleocapsid protein's role as a shuttle protein responsible for transporting viral genome across membranes [35]. This method provides us with a basis to know which protein family is a prominent member. which can be targeted for rendering a virus incapable of affecting a human host.

Table 4. Protein profiles of the 3 strains

	SARS-COV-2	**SARS**	**Bat**
Genome length	29868	29706	29059
Total amino acids	9956	9902	9686
Total proteins	747	728	699
Functional proteins	89	84	76
Number of family classes	41	44	42

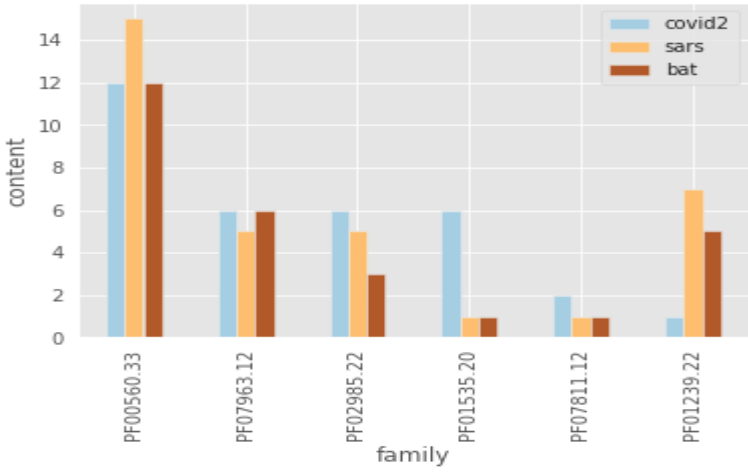

Fig. 7. Pfam content in each genome

Though Bi-LSTM is superior to RNNs in this application as Bi-LSTM could retain the "memory" of outcomes of the previous layers, but they are inferior to GRU due to its higher complexity and computer cost. GRUs are simpler yet retain retains the ability to train for long-term dependency, easier to modify and can be trained faster than LSTMs [26, 36] with same number of trainable parameters. Moreover GRUs show better performance for smaller and less frequent dataset [37, 38]. In our experiment we show that GRU outperforms Bi-LSTM in terms of accuracy of multi-class protein classification and the model loss criterion. Exploration of Temporal Convolutional Network (TCN) is another avenue to successfully model the protein sequences [39] where TCN used significantly lesser memory than LSTM and GRUs to store partial results. Bai et al. [40] showed that TCN outperformed recurrent architectures in modeling gene sequence for large genetic dataset. In our research we focus on GRU as the methodology for classification and compared the performance with that of Bi-LSTM model. We will explore TCN in our application as future works.

6. Conclusion

The methodology and the proposed classification model discussed above are based on the complete genome data of an organism. The initial genome sequence was released by the Chinese authorities in mid-January, 2020 as detected in the first patients. This followed by other institutions releasing the genomic sequences starting late January 2020. The outcomes of our study demonstrate that the CoV-2 virus genome structure is more similar to the bat virus than the SARS virus. The results show that there is 82.42 % match between CoV-2 and bat virus genome structures. In addition, we have found that PF00560.3 is the protein family accession that is prominent in all the virus structures. Our GRU model predicts PF00560.3 as the prominent protein family for CoV-2 with an average probability of 76% computed from the probability distribution generated by the Softmax function output against the protein families. Our finding infers that CoV-2 proteins turns out to be functionally similar to the proteins found in previous strains as observed in prior research by Reed et al. [25].

In this study we self-imposed few limitations to restrict the GRU model to classify protein families with only top 1000 protein families due to considerable computational resources required to train and test the model. Nevertheless, the results are notable and provides insights to understand coronavirus and its protein classification. The results can further be employed in medical research, considering the functionalities of PF00560.3 and how they are responsible for affecting a human host. In addition, our results identify the protein family that can be used to develop counterfeit measures against the Coronavirus.

In future work we would suggest to carry out a comparative study with the annotated sequences to power up our existing model. We would work to extend our classifier and improve its accuracy with optimized hyperparameters to include all protein families. We also intend to study the performance of our model using one versus rest (binary) classifier for the prediction of the protein family in our future work. Additional models like TCN with hyperparameter tuning will be explored in future in addition to GRU and Bi-LSTM model presented in this research.

Acknowledgments

We acknowledge Ramya Chimata Venkatakrishnan and Venkat Nihaal Akula for their help in this project. Any remaining errors are authors' responsibility.

References

1. Lu R., Zhao X., Li J., Niu P., Yang B., Wu H., et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet, 395(10224), 565-574 (2020). DOI: https://doi.org/10.1016/S0140-6736(20)30251-8.
2. Guo Y.-R., Cao Q.-D., Hong Z.-S., Tan Y.-Y., Chen S.-D., Jin H.-J., et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status. Military Medical Research, 7(1), 1-10 (2020). DOI: https://doi.org/10.1186/s40779-020-00240-0.
3. Ruan Y., Wei C. L., Ling A. E., Vega V. B., Thoreau H., Thoe S. Y. S., et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. The Lancet, 361(9371), 1779-1785 (2003). DOI: https://doi.org/10.1016/S0140-6736(03)13414-9.
4. Fehr A. R., Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Coronaviruses. Methods of Molecular Biology, 1282, 1-23 (2015). DOI: 10.1007/978-1-4939-2438-7_1.
5. Wu F., Zhao S., Yu B., Chen Y.-M., Wang W., Song Z.-G., et al. A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265-269 (2020).

6. Zhou P., Yang X.-L., Wang X.-G., Hu B., Zhang L., Zhang W., et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 579(7798), 270-273 (2020).
7. De Wit E., Van Doremalen N., Falzarano D., Munster V. SARS and MERS: recent insights into emerging coronaviruses. Nature Reviews Microbiology, 14(8), 523-534 (2016).
8. Wu A., Peng Y., Huang B., Ding X., Wang X., Niu P., et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell host & Microbe, 27(13), 325-328 (2020). DOI: https://doi.org/10.1016/j.chom.2020.02.001.
9. Angeletti S., Benvenuto D., Bianchi M., Giovanetti M., Pascarella S., Ciccozzi M. COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. Journal of medical virology, 92(6), 584-588 (2020). DOI: https://doi.org/10.1002/jmv.25719.
10. Tang X., Wu C., Li X., Song Y., Yao X., Wu X., et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review, 7(6), 1012–1023 (2020). https://doi.org/10.1093/nsr/nwaa036.
11. Zhu N., Zhang D., Wang W., Li X., Yang B., Song J., et al. A novel coronavirus from patients with pneumonia in China, 2019. New England Journal of Medicine, 382(8), 727-733 (2020). DOI: 10.1056/NEJMoa2001017.
12. Bileschi M. L., Belanger D., Bryant D. H., Sanderson T., Carter B., Sculley D., et al. Using deep learning to annotate the protein universe. bioRxiv, 1-28. (2019). DOI: https://doi.org/10.1101/626507.
13. LeCun Y., Bengio Y., Hinton G. Deep learning. Nature, 521(7553), 436-444 (2015).
14. Bateman A., Coin L., Durbin R., Finn R. D., Hollich V., Griffiths-Jones S., et al. The Pfam protein families database. Nucleic acids research, 32(suppl_1), D138-D141 (2004). DOI: https://doi.org/10.1093/nar/gkh121.
15. D'Agaro E. Artificial intelligence used in genome analysis studies. The EuroBiotech Journal, 2(2), 78-88 (2018). DOI: https://doi.org/10.2478/ebtj-2018-0012.
16. Vijay R. Protein Sequence Classification: A case study on Pfam dataset to classify protein families. https://towardsdatascience.com/protein-sequence-classification-99c80d0ad2df. Last accessed 2019/09/02.
17. Hu H., Li Z., Elofsson A., Xie S. A Bi-LSTM based ensemble algorithm for prediction of protein secondary structure. Applied Sciences, 9(17), 3538 (2019).
18. Jurtz V. I., Johansen A. R., Nielsen M., Almagro Armenteros J. J., Nielsen H., Sønderby C. K., et al. An introduction to deep learning on biological sequence data: examples and solutions. Bioinformatics, 33(22), 3685-3690 (2017). DOI: https://doi-org.ezaccess.libraries.psu.edu/10.1093/bioinformatics/btx531.
19. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J. Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410 (1990). DOI: https://doi.org/ 10.1016/S0022-2836(05)80360-2.
20. Ye J., McGinnis S., Madden T. L. BLAST: improvements for better sequence analysis. Nucleic acids research, 34(suppl_2), W6-W9 (2006). DOI: https://doi.org/10.1093/nar/gkl164.
21. McGinnis S., Madden T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic acids research, 32(suppl_2), W20-W25 (2004). DOI: https://doi.org/10.1093/nar/gkh435.
22. Yuan J., Hon C.-C., Li Y., Wang D., Xu G., Zhang H., et al. Intraspecies diversity of SARS-like coronaviruses in Rhinolophus sinicus and its implications for the origin of SARS coronaviruses in humans. Journal of general virology, 91(4), 1058-1062 (2010). DOI: https://doi.org/10.1099/vir.0.016378-0.
23. Wheeler D. L., Barrett T., Benson D. A., Bryant S. H., Canese K., Chetvernin V., et al. Database resources of the national center for biotechnology information. Nucleic acids research, 36(suppl_1), D13-D21 (2007). DOI: https://doi.org/10.1093/nar/gkm1000
24. Song S., Huang H., Ruan T. Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools Applications, 78(1), 857-875 (2019). DOI: https://doi.org/10.1007/s11042-018-5749-3
25. Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. in EMNLP, Association for Computational Linguistics, pp. 1724-1734.
26. Sherstinsky A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena, 404(132306), 1-28 (2020). DOI: https://doi.org/10.1016/j.physd.2019.132306.
27. Zulqarnain M., Ghazali R., Ghouse M. G., Mushtaq M. F. Efficient processing of GRU based on word embedding for text classification. International Journal on Informatics Visualization, 3(4), 377-383 (2019). DOI: 10.30630/joiv.3.4.289

28. Lee T. K., Nguyen T. Protein family classification with neural networks. Stanford University, pp. 1-9 (2016).
29. Le N. Q. K., Yapp E. K. Y., Nagasundaram N., Chua M. C. H., Yeh H.-Y. J. C. Computational identification of vesicular transport proteins from sequences using deep gated recurrent units architecture. Computational and Structural Biotechnology, 17, 1245-1254 (2019). DOI: https://doi.org/10.1016/j.csbj.2019.09.005.
30. Pfeiffenberger E., Bates P. A. Predicting improved protein conformations with a temporal deep recurrent neural network. PLos One, 13(9), e0202652 (2018). DOI: https://doi.org/10.1371/journal.pone.0202652.
31. Le N. Q. K., Yapp E. K. Y., Yeh H.-Y. ET-GRU: using multi-layer gated recurrent units to identify electron transport proteins. BMC Bioinformatics, 20(1), 1-12 (2019).
32. Zhao M., Wang H., Guo J., Liu D., Xie C., Liu Q., et al. Construction of an industrial knowledge graph for unstructured chinese text learning. Applied Sciences, 9(13), 2720. (2019). DOI: 10.3390/app9132720.
33. Needleman S. B., Wunsch C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, 48, 443-153 (1970). DOI: https://doi.org/10.1016/0022-2836(70)90057-4.
34. Reed M. L., Howell G., Harrison S. M., Spencer K.-A., Hiscox J. A. Characterization of the nuclear export signal in the coronavirus infectious bronchitis virus nucleocapsid protein. Journal of virology, 81(8), 4298-4304 (2007). DOI: https://doi.org/10.1128/JVI.02239-06.
35. Timani K. A., Liao Q., Ye L., Zeng Y., Liu J., Zheng Y., et al. Nuclear/nucleolar localization properties of C-terminal nucleocapsid protein of SARS coronavirus. Virus research, 114(1-2), 23-34 (2005). DOI: https://doi.org/10.1016/j.virusres.2005.05.007
36. Gers F. A., Schmidhuber J., Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation, 12(10) 2451-2471 (2000). DOI: https://doi.org/10.1162/089976600300015015.
37. Chung J., Gulcehre C., Cho K., Bengio Y. (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS Workshop on Deep Learning. DOI: http://arxiv.org/abs/1412.3555.
38. Gruber N., Jockisch A. Are GRU cells more specific and LSTM cells more sensitive in motive classification of text? Frontiers in Artificial Intelligence, 3, 1-6 (2020). DOI: https://doi.org/10.3389/frai.2020.00040.
39. Kim H. Y., Kim D. Prediction of mutation effects using a deep temporal convolutional network. Bioinformatics, 36(7), 2047-2052 (2020). DOI: https://doi.org/10.1093/bioinformatics/btz873.

Aims and Objectives

Published online by ICS two times a year, Journal of Digital Science (JDS) is an international peer-reviewed journal which aims at the latest ideas, innovations, trends, experiences and concerns in the field of digital science covering all areas of the scholarly literature of the sciences, social sciences. The main topics currently covered include: Artificial Intelligence Research; Digital Economics, Education, Engineering, Finance, Health Care.

The main goal of the journal is the effective dissemination of original incites/results generated by the human brain and presented/reflected in articles using modern information/digital technology.

Editorial Board

Contact information

Website: https://ics.events

Email: conf@ics.events