ARAB JOURNAL OF BASIC AND APPLIED SCIENCES
2023, VOL. 30, NO. 1, 103–121
https://doi.org/10.1080/25765299.2023.2188677

University of Bahrain
Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Mutation prediction and phylogenetic analysis of SARS-CoV2 protein sequences using LSTM based encoder-decoder model

Sweeti Sah[a,b], B. Surendiran[a,b], R. Dhanalakshmi[a,c] and Sachi Nandan Mohanty[a,d]

[a]Department of Computer Science & Engineering, National Institute of Technology Puducherry, Karaikal, India; [b]National Institute of Technology Puducherry, Karaikal, India; [c]IIIT, Tiruchirappalli, Trichy, India; [d]School of Computer Science & Engineering (SCOPE), VIT-AP University, Amaravati, Andhra Pradesh, India

ABSTRACT

The ongoing evolution and mutation of SARS-CoV2 pose a significant challenge to the development of effective medication, as genetic changes can render previously developed drugs ineffective. To address this issue, researchers are exploring various strategies to predict and assess the emergence of novel SARS-CoV2 strains through phylogenetic analysis and mutation prediction. In recent years, deep learning approaches have been applied to studying viruses, including SARS-CoV2, to improve our understanding of virus evolution, structure, categorization, and prediction. In this study, a novel deep learning approach is proposed to predict and assess SARS-CoV2 protein sequences. Specifically, Long Short-Term Memory (LSTM) is utilized to predict protein sequences from aligned input sequences, with a bioinformatics tool used to detect mutations. The deep learning model proposed in this study exhibits high accuracy in predicting several key SARS-CoV2 protein sequences, including spike, replicase, putative, ORF1a, and nucleocapsid. The study uses genome sequencing data from the National Center for Biotechnology Information (NCBI) and demonstrates a 98% accuracy in predicting genomic sequences, with minimal changes observed in protein sequences. This study represents a significant improvement over previous research, which has focused only on predicting mutations in viral RNA sequences using datasets from other viruses.

ARTICLE HISTORY
Received 6 October 2022
Revised 20 February 2023
Accepted 3 March 2023

KEYWORDS
LSTM; prediction; Seq2Seq; genomic sequence; protein; SARS-CoV2; alignment; phylogenetic tree

1. Introduction

Adenine (A), thymine (T), cytosine (C), and guanine (G) are the four nitrogen-containing nucleobases that make up all nucleotides (G). The RNA sequence differs from the DNA sequence because it has a more significant mutation and is more stable (Mohamed, Sayed, Salah, & Houssein, 2021). SARS-CoV-2 is scattering rapidly due to the inaccuracy of current recognition technologies (Lopez-Rincon et al., 2021). SARS-CoV-2, on the other hand, is a typical RNA virus that generates new mutations in a Coronavirus replication cycle, including 10-4 nucleotide substitutions per year is the usual evolutionary rate each year per site (Lu et al., 2020). SARS-CoV2 belongs to the Coronaviridae family (Whata & Chimedza, 2021), and its identification can be challenging due to mutations. So, this paper has explored the concepts of detecting the mutation and prediction of sequences using the deep learning method. Having access to current virus mutations and prior evolution could help researchers better understand virus evolution dynamics and predict future viruses and diseases (Shendure & Ji, 2008).

In human disease genetics, the prediction of genetic mutations is a hot topic (Stranger & Dermitzakis, 2006). Knowing about current virus generations and their prior evolution could serve to understand the dynamics of virus evolution and forecast future viruses and diseases (Shendure & Ji, 2008). The ancestral sequence of these species is inferred via phylogenetic analysis, which determines the evolutionary relationship between them. These evolutionary connections between RNA sequences can help anticipate which lines may have the same function (Xu et al., 2015).

This paper presents several significant contributions in the field of bioinformatics. First, a novel method is proposed for the alignment of protein sequences to identify mutations and assess the similarity between genomic sequences. This technique employs advanced algorithms for sequence alignment and statistical analysis, enabling accurate and

CONTACT Sachi Nandan Mohanty sachinandan09@gmail.com School of Computer Science & Engineering (SCOPE), VIT-AP University, Amaravati, Andhra Pradesh, India.

reliable comparisons of protein sequences. Second, an evolution tree is generated for the protein sequences of SARS-CoV2, providing insight into the relationships and origins of different strains of the virus. Third, a Long Short-Term Memory (LSTM) based Encoder-Decoder deep learning model is developed to predict mutations in protein sequences of SARS-CoV2. This model utilizes machine learning algorithms to analyze large datasets of protein sequences and associated mutation data, enabling accurate predictions of specific mutations in the viral genome. Finally, it also presents a method for predicting nucleotide changes and identifying new strains of the virus in the new generation. Overall, these contributions represent significant advances in the study of SARS-CoV2 and provide valuable tools and techniques for understanding the virus's evolution, pathogenicity, and potential for developing new treatments and vaccines. Hence, the approach taken in this study provides a more comprehensive analysis of the mutations present in SARS-CoV2 protein sequences and has the potential to improve our ability to predict and respond to emerging strains of the virus.

There are several tools available for aligning protein sequences, including:

Clustal Omega: This is a popular online tool for multiple sequence alignment of proteins. You can input up to 500 sequences in FASTA or Clustal format and choose different options for alignment parameters. The output can be visualized as an alignment or a tree (Sievers & Higgins, 2018).

MUSCLE: This is another online tool for protein sequence alignment. It allows you to input up to 500 sequences and provides options for alignment parameters, such as the gap opening penalty and the gap extension penalty (Edgar, 2004).

T-Coffee: This tool provides a variety of alignment methods and allows you to input multiple sequence formats, including FASTA, EMBL, and UniProt. T-Coffee also allows you to visualize the alignment output in a variety of ways (Taly et al., 2011).

BioEdit: It is a popular desktop software tool for sequence alignment, visualization, and analysis. It is widely used by researchers and has many useful features for working with DNA, RNA, and protein sequences. One of the key features of BioEdit is its ability to align multiple sequences using a variety of algorithms, including ClustalW, T-Coffee, and MUSCLE. The software also allows for manual editing of alignments, which can be useful for fine-tuning the alignment or correcting errors. In addition to alignment, BioEdit can be used for a variety of other tasks, such as sequence annotation, primer design, and restriction enzyme analysis. The software also includes visualization tools, such as the ability to generate graphical representations of alignments or sequence features (Hall, 1999; Tomita, Mori, & Mochizuki, 2015; Carvalho, Fischer, & Chen, 2009).

Alternatively, we can use bioinformatics software packages, such as MEGA, which allow you to align multiple sequences, generate phylogenetic trees, and perform other analyses. These software packages are typically more powerful and flexible than online tools but require more expertise to use effectively.

There have been various traditional machine learning approaches like Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), and various deep learning approaches for predicting the sequences like Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), Long Short Term Memory (LSTM). The advantage of using the deep learning approach is that it permits variable length sequences as input and output. Long Short Term Memory has been extensively used in the literature for predicting the genomic sequences of other viruses. Because LSTM is capable of capturing the longer sequences with several gating mechanisms (Zhou et al., 2023).

Predicting mutations in protein sequences is an important task in the field of bioinformatics, and there are several tools and techniques available to do so. One approach is to use in silico methods to predict the impact of mutations on protein structure and function. This can be done using software programs that analyze the effects of amino acid changes on the physical and chemical properties of the protein, such as its stability, solubility, and interactions with other molecules. Another approach is to use machine learning algorithms to predict the likelihood of specific mutations occurring in a given protein sequence. Machine learning models can be trained on large datasets of protein sequences and associated mutation data to learn patterns and make predictions about the likelihood of specific mutations occurring. Overall, predicting mutations in protein sequences is an important area of research for understanding the evolution and pathogenicity of SARS-CoV-2, as well as for developing new treatments and vaccines (Kumar, Stecher, & Tamura, 2016).

The following is the structure of the entire paper: The introduction to the research work is included in Section 1. Section 2 contains a review of previous research. Section 3 details the data sources and methods used in the prediction. The intended work's outcome is shown in Section 4. Finally, section 5 summarizes the work that can be done and its future scope.

2. Literature review

SARS-CoV-2 is an RNA virus, and like all RNA viruses, it has a high mutation rate. Mutations are changes in the genetic material (in this case, RNA) of the virus. Mutations can be beneficial, harmful, or neutral to the virus, depending on their effects on the virus's survival and ability to replicate. There have been many mutations identified in SARS-CoV-2 since the start of the pandemic. Some of these mutations are more significant than others and can affect the behavior of the virus. One particular mutation, known as the D614G mutation, has been associated with increased transmissibility of the virus. Other mutations have been identified in the spike protein of the virus, which is the protein that allows the virus to enter and infect human cells. Some of these mutations may make the virus more infectious or more resistant to antibodies generated by vaccination or previous infection. It's important to note that not all mutations are necessarily a cause for concern. Many mutations may not affect the behavior of the virus or may even weaken the virus. However, monitoring mutations is an important part of understanding how the virus is evolving and how it may respond to vaccines and treatments. That's why ongoing genomic surveillance is crucial in tracking the spread and evolution of SARS-CoV-2 (Rambaut et al., 2020).

The mutation rate of the complete genome sequence of SARS-CoV-2 has been investigated using patient datasets from various countries. Based on the collected data, specific nucleotide and codon mutations have been identified. The mutation rate has been divided into four groups according to the dataset size: China, Australia, the United States, and the rest of the world. Although codons have a lower mutation rate than nucleotides, a substantial number of thymine (T) and adenine (A) nucleotides have been found to change to other nucleotides in all locations. The Long Short-Term Memory (LSTM) model has been used to predict the nucleotide mutation rate of the 400th patient. The mutation rate increases by 0.1 percent when nucleotides change from T to C and G, C to G, and G to T, whereas changing T to A and A to C lowers the score by 0.1 percent. The study explores how COVID-19 genomic sequences can be utilized to extract meaningful information using artificial intelligence methods. Sequential Pattern Mining (SPM) is first applied to a corpus of machine-readable COVID-19 genome sequences to determine whether any significant hidden patterns, such as recurrent patterns of nucleotide bases and their interactions, may be discovered. Sequence predictions are then applied to the corpus to determine whether nucleotide bases can be anticipated from earlier ones. Finally, an algorithm is developed for genome sequence mutation analysis to identify regions in genome sequences where nucleotide bases change and to determine the mutation rate. The results demonstrate that by utilizing SPM and mutation analysis techniques, it is possible to detect intriguing trends in the COVID-19 genomic sequences, allowing for the evaluation of the evolution and variability of COVID-19 strains (Pathan, Biswas, & Khandaker, 2020; Nawaz, Fournier-Viger, Shojaee, & Fujita, 2021).

Few Researchers used the seq2seq LSTM neural network to predict next-generation sequences by using the method while treating the sequences as textual data (Mohamed et al., 2021). As a result of using single hot vectors as input, the model retains the important information position of each nucleotide in the sequences. Two RNA virus sequencing datasets were used to test the proposed model, and the findings were promising. The results show how the LSTM neural network for DNA and RNA sequences can be used to handle a variety of bioinformatics sequencing difficulties (Chen, Gao, Wang, & Wei, 2021). examines the mechanism, frequency, and ratio of mutations in the S protein, which is a frequent target of the majority of COVID-19 vaccines and antibody treatments. 56 antibody constructions were also generated, and their 2D and 3D properties were studied. Additionally, it is anticipated that mutations will change the binding free energies (BFE) of S protein and antibody or ACE2 complexes. The majority of the 462 mutations on the receptor-binding domain (RBD) degrade the binding of S protein and antibodies, jeopardizing the effectiveness and dependability of antibody treatments and vaccinations, according to research that combines genetics, biophysics, deep learning, and algebraic topology (Nguyen et al., 2021). Utilizing deep learning approaches, this study describes and analyses genetic changes in SARS-coding CoV-2 areas, as well as their predicted effects on protein secondary structure and solvent accessibility. The predictions indicate that the highly publicized mutation D614G in the viral spike protein is unlikely to affect the protein's secondary structure or relative solvent availability. Based on 6324 viral genome sequences, the author created a mutational spreadsheet dataset to support research into SARS-CoV-2 from a variety of angles, particularly in tracing the virus's evolution and global distribution. The results also demonstrate that E, M, ORF6, ORF7a, ORF7b, and ORF10 are the most stable coding genes, suggesting that these genes may be used to create vaccines and treatments.

The most recent COVID-19 pandemic is currently raging, with new strains including surprising changes. Understanding how to predict virus

alterations has important implications for developing vaccines and medications, and prevention strategies. Because the number of reported changes in SARS-CoV-2 is currently restricted, creating a prediction model employing virus data with many mutations, such as the influenza A virus, would be advantageous and straightforward. The likelihood mutation sites and changed amino acids in hemagglutinins from the Eurasia H1 influenza A virus were predicted using a neural network with a feedforward backpropagation algorithm in this study (Yan & Wu, 2021). The purpose of the study is to use one of the most comprehensive data sets available, which includes 506,768 SARS-CoV-2 genome sequences, to follow fast-spreading RBD mutations in pandemic-affected countries and investigate their evolutionary tendencies around the world. There were 6945 unique single mutations found on the S protein, with 1024 of them occurring on the RBD. 100 of the 651 non-degenerate variants on the RBD were detected more than 28 times in the database and deemed significant protein sequence alterations. Also, it showed that in addition to the N501Y, E484K, and K417N modifications in the UK, South Africa, and Brazil variations, L452R, E484Q mutations in India, S477N, N439K, S477R, and N501T variations in 31 disease outbreak countries in the last few months, N439K, S477R, S477N, and N501T mutations (Wang, Chen, Gao, & Wei, 2021).

There are many deep learning models that can be used for predicting amino acid sequences. Some of the latest models are, Alphafold which is developed by DeepMind, Alphafold uses deep learning to predict the 3D structure of proteins. It won the 2020 CASP14 competition by accurately predicting the structures of 25 out of 43 proteins. RoseTTAFold, which is developed by the University of Washington, RoseTTAFold uses a combination of deep learning and template-based modeling to predict the 3D structure of proteins. It outperformed other methods in the CASP14 competition. TAPE which is developed by the University of California, Berkeley, TAPE (The TAProot Ensemble) is a deep learning model that can predict various protein properties, including secondary structure, solvent accessibility, contact prediction, and remote homology detection. ProGenc, which is developed by Stanford University, ProGen is a deep learning model that can predict the amino acid sequence of a protein from its 3D structure. UniRep which is developed by Harvard University, UniRep is a deep learning model that can encode protein sequences into fixed-length vectors that can be used for various downstream tasks, such as protein function prediction and protein-protein interaction prediction. These models are constantly being improved upon and new models are also being developed, so it's always worth keeping up to date with the latest research in the field.

3. Data source and methods

3.1. Data source

Predicting SARS-CoV-2 mutations is a complex task that requires expertise in both bioinformatics and deep learning. LSTM encoder-decoder models have been used in many natural language processing tasks, but they can also be applied to sequence prediction tasks, such as SARS-CoV-2 mutation prediction. Our evaluation collected the dataset from the National Center for Biotechnology Information (NCBI) (National Center for Biotechnology Information \(NCBI\) Bethesda \(MD\), 1988). Total 250 SARS-CoV2 variants were considered (Sah, Surendiran, & Dhanalakshmi, 2023). This is the world's largest dataset repository for genomic sequences. The information gathered pertains to all protein sequences. Our dataset is in FASTA format. The experimental setup for predicting the mutation rate of sequences is shown in Table 1.

Table 1 contains information about the experimental requirements for the proposed work.

An LSTM-based model can be trained on a large dataset of protein sequences and their corresponding mutation information to learn patterns and relationships between the sequences and mutations. The model can then be used to predict mutations in new protein sequences based on those patterns and relationships. In the case of SARS-CoV-2, an LSTM-based model can be trained on a dataset of protein sequences from different strains of the virus and their associated mutation information. The model can learn how different mutations affect the structure and function of the viral proteins and use that knowledge to predict the effects of new mutations. The input to the model is a sequence of amino acids that make up the protein, and the output is the predicted mutation(s) and their effects on the protein. The model uses the previous state of the LSTM to encode the sequence and then decodes it to generate the prediction. The LSTM model can be a

Table 1. Experimental setup for the proposed model.

Dataset Used	Genomic Sequence
Dataset Format	FASTA
Deep Learning Model	LSTM (For predicting sequences)
Language	Python
Software	Colab
Training Data	80%
Validation Data	20%
Activation Function	Softmax
Epoch	50
Batch Size	10
Loss	Categorial Cross entropy
Optimizer	Adam
Bio informatics Tool	Bioedit (For analyzing sequences)

Table 2. Contribution summary.

Contribution	Approach
Collected Amino Acid Sequences or Protein Sequences from several SARS-CoV2 nucleotide sequences that we classified using machine learning approaches. Alignment of amino acid sequences done.	Bioinformatics Approach (Step 1)
Predicted mutation as no literature explored the prediction of protein sequences of SARS-CoV2.	Deep Learning Approach (Step 2)

Table 3. Total length of protein sequences before and after alignment.

Protein	Total Sequences Considered	After Alignment Length
Nucleocapsid protein	32	422
ORF1a	65	4377
Putative Spike Glycoprotein	26	1255
Replicase	45	4376
Spike Glycoprotein	101	1255

powerful tool for predicting mutations in protein sequences, but it is important to note that the accuracy of the predictions depends on the quality and size of the training dataset, as well as the features and parameters of the model itself.

Amino acids are a group of 20 chemicals that make up proteins. Proteins are made up of polypeptides, long chains of amino acids. The amino acid chain sequence causes the polypeptide to fold into a physiologically active form. Protein amino acid sequences are stored in the genes (Smith, 2019; Sah, Surendiran, Dhanalakshmi, & Kamerkar, 2021). Annexure 1 shows the protein names considered for the experimentation, showing the entire common amino acid sequences and sequence length for each protein after alignment. The amino acid sequences are collected from NCBI (National Center for Biotechnology Information \(NCBI\) Bethesda \(MD\), 1988).

3.2. Proposed LSTM model

Long short-term memory (LSTM) is a variant of RNNs (Hochreiter & Schmidhuber, 1997) that can learn long-term dependencies and is specifically designed to avoid the problem of long-term dependencies. In the context of protein sequence prediction, LSTMs have been used to predict the secondary structure of proteins, as well as the binding affinity between proteins and ligands. LSTMs can also be used to predict the sequence of a protein from its genetic sequence, which is an important step in drug design and other bioinformatics applications. LSTMs work by passing information through a series of "gates" that control the flow of information through the network. These gates allow the LSTM to selectively remember or forget previous inputs, which enables it to maintain a long-term memory of the input sequence. The output of the LSTM is then used to make a prediction about the next item in the sequence. To train the LSTM model, the protein sequence data is converted to one-hot encoded format using the np.eye(n_classes) function, which creates an identity matrix with n_classes rows and columns. Each row of the identity matrix corresponds to an amino acid, and each column corresponds to a position in the protein sequence. The seq_data_one_hot variable is a 3D numpy array with shape (n_samples, max_seq_len, n_classes). The LSTM model is defined using the Keras Sequential model API, with one LSTM layer and one dense output layer. The model is compiled with the appropriate loss function and optimizer.

The protein sequence data is split into training and validation sets, and the model is trained using the fit method of the Keras Sequential model API. Table 2, shows the flow and contribution for the proposed work.

Protein sequences can be fed into an LSTM model for training using Python and the Keras deep learning library. A LSTM unit has a cell/node, an input gate, an output gate, and a forget gate at its base. The node considers values during particular time intervals, while the input/output gates control the information flow (Koumakis, 2020). Long Short-Term Memory (LSTM) model is proposed to predict the amino acid sequences or protein sequences of the virus. The proposed work consists of a few steps. In the initial step, the protein sequence is preprocessed. Before alignment total sequences considered were 149, each length varying from 5k to 6k. After alignment, the sequence length considered as shown in Table 3.

Now, the LSTM model applies one hot encoding representation of the sequences to the same-length input. The one-hot encoded vector is added by each LSTM cell to the hidden state and cell state vectors. The third phase of the encoder output is the cell state and hidden state concealed values vectors. With the exception of the first cell, which receives its cell and hidden states directly from the encoder, each subsequent cell now derives its cell and hidden states from the one before it. The probability distribution of the word at position t in the succeeding

generation sequence is predicted using the dense layer. A phylogenetic tree is a branching diagram that represents the evolutionary relationships among a set of organisms or sequences, based on the similarities and differences in their genetic or protein sequences. In the case of SARS-CoV-2, a phylogenetic tree can be constructed using the amino acid sequences of the virus. The tree can help to visualize the evolutionary history of SARS-CoV-2, and can be used to identify the origin of the virus, its transmission patterns, and the emergence of new variants. The tree is typically constructed using bioinformatics software that can align the amino acid sequences, calculate the genetic distances between them, and infer the branching patterns. One common software used to construct phylogenetic trees is MEGA (Molecular Evolutionary Genetics Analysis), which can handle large datasets and provide various options for phylogenetic analysis, including maximum likelihood, neighbor-joining, and Bayesian inference. Other software tools used for phylogenetic analysis include RAxML, PhyML, and BEAST. The resulting tree can be visualized using software such as FigTree or iTOL (Interactive Tree of Life), which allows for further customization and annotation of the tree. The phylogenetic tree can provide valuable insights into the evolution and diversity of SARS-CoV-2, and can inform public health measures and vaccine development strategies.

Workflow:

Step 1: Import necessary Libraries
Step 2: Load Amino Acid Sequences
Step 3: For each sequence, map unique chars to integer by creating dictionary
Step 4: Prepare the genomic dataset of input to output pairs encoded as integer
Step 5: Reshape the data
Step 6: Apply one hot encoding
Step 7: Define LSTM model
Step 8: Define the checkpoint
Step 9: Fit the model
Step 9: Load the network weights
Step 10: Compute Accuracy

Figure 1, shows the workflow for the proposed work for predicting the mutations in the protein sequence.

The mutations in SARS-CoV-2 can have a significant impact on pathogenicity, diagnostics, therapeutics, and vaccines. Here are some of the ways in which mutations can impact each of these areas (Centers for Disease Control & Prevention, 2021; World Health Organization, 2021; Korber et al., 2020; Lauring & Hodcroft, 2021):

i. **Pathogenicity:** Mutations in SARS-CoV-2 can affect how the virus interacts with the host cells, leading to changes in the severity of the disease. For example, some mutations have been associated with increased transmission and more severe disease, while others have been associated with decreased virulence. Mutations in the spike protein can affect the virus's ability to bind to the ACE2 receptor on host cells, which is a key step in the viral infection process.
ii. **Diagnostics:** Mutations in SARS-CoV-2 can impact diagnostic tests, particularly those that rely on detecting viral RNA. For example, some

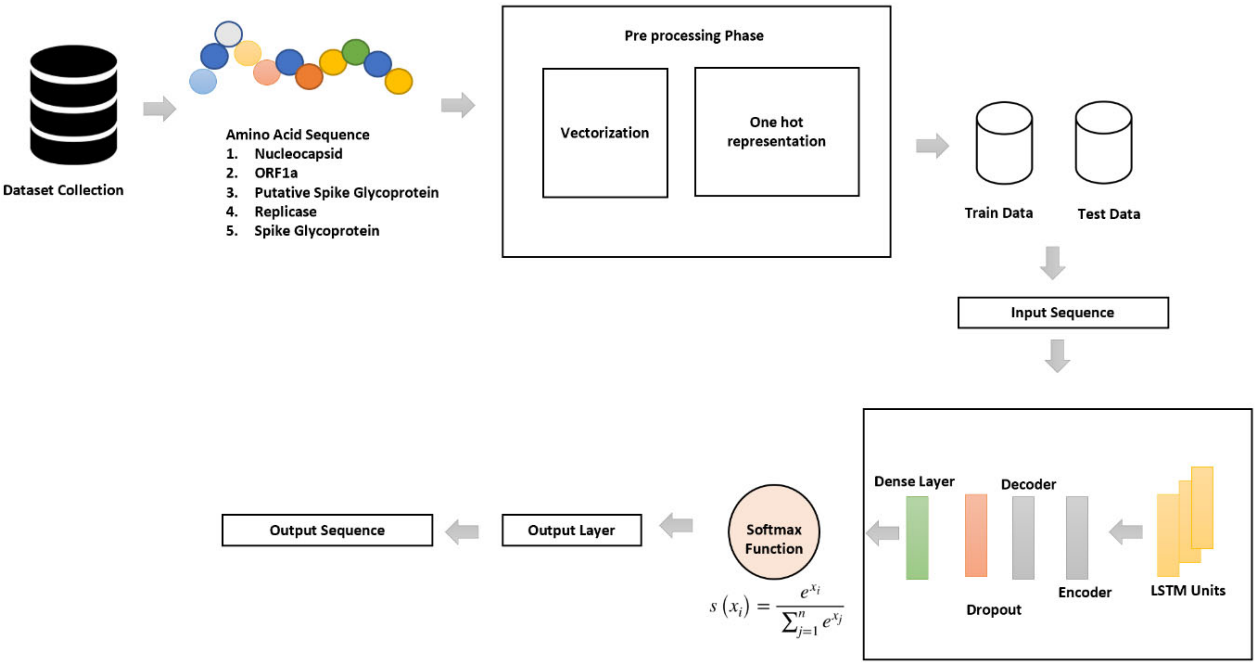

Figure 1. Proposed model for predicting the sequences.

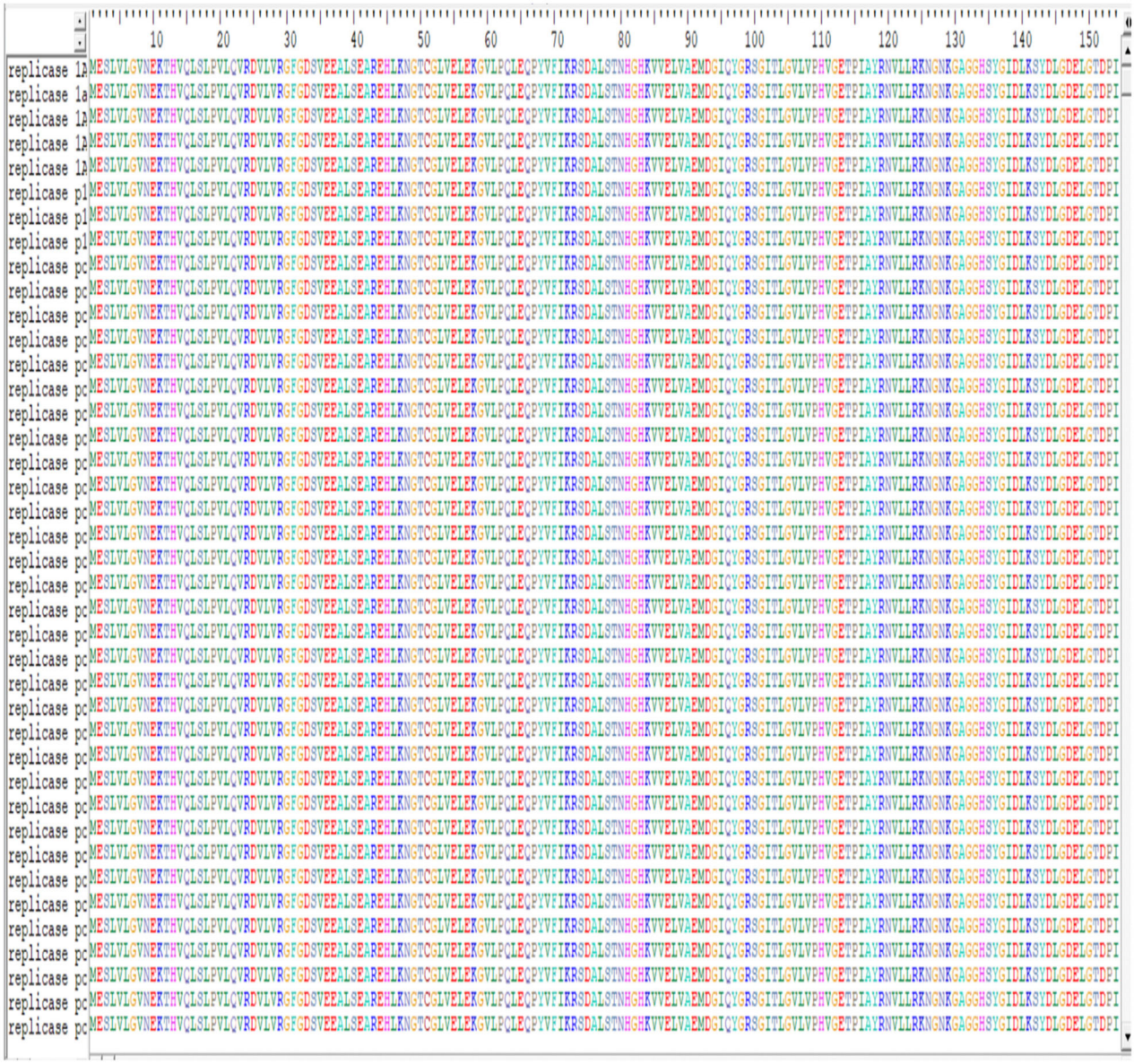

Figure 2. Alignment of replicase proteins.

mutations can cause false negative results in PCR tests, which can lead to incorrect diagnoses and potentially contribute to the spread of the virus. New strains of the virus that carry multiple mutations may require updates to existing diagnostic tests to ensure their accuracy.

iii. **Therapeutics:** Mutations in SARS-CoV-2 can impact the effectiveness of therapeutic treatments. For example, some mutations in the spike protein can affect the binding of neutralizing antibodies, making certain treatments less effective. The emergence of new strains of the virus can also impact the effectiveness of existing treatments and require the development of new therapies.

iv. **Vaccines:** Mutations in SARS-CoV-2 can impact the effectiveness of vaccines. For example, mutations in the spike protein can affect the ability of the immune system to recognize and neutralize the virus. If a mutation occurs in a region of the virus that is targeted by a vaccine, it can reduce the vaccine's effectiveness. The emergence of new strains of the virus may require the development of updated or new vaccines to ensure their effectiveness.

While complementary and alternative medicinal plants, such as 6-shogaol, have been shown to have potential therapeutic properties, their effectiveness as a treatment for an evolving virus like SARS-CoV-2 is uncertain, and they should not be used as a replacement for conventional treatments. 6-shogaol is a natural compound found in ginger and has been studied for its potential medicinal properties, including antiviral activity. Some studies have suggested that 6-shogaol may have activity against various viruses, including influenza, herpes simplex virus, and respiratory syncytial virus. However, there is

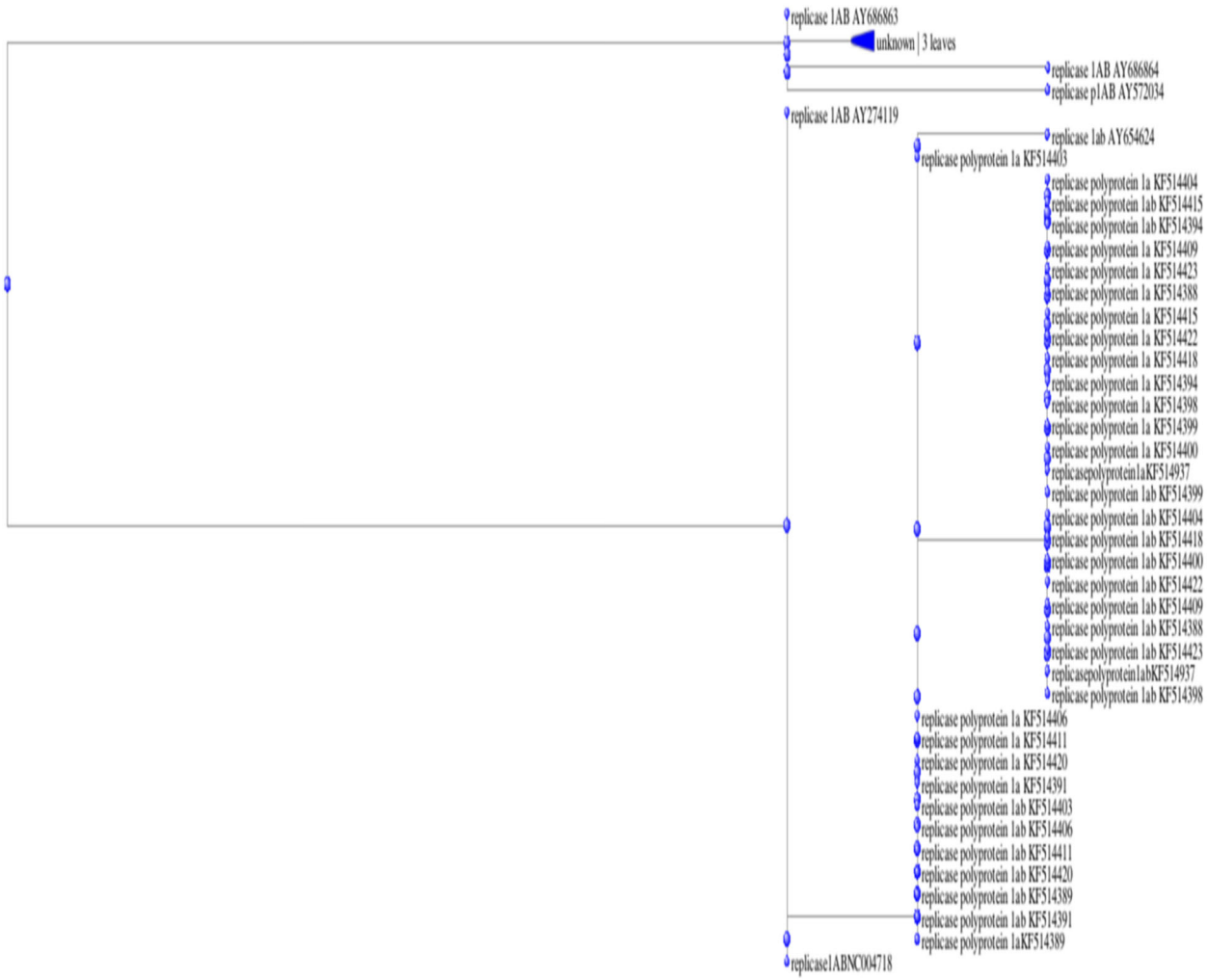

Figure 3. Phylogenetic tree of replicase proteins.

currently no clinical evidence to support the use of 6-shogaol as a treatment for COVID-19, and more research is needed to determine its safety and effectiveness.

It is important to note that conventional treatments, such as vaccines, antiviral drugs, and supportive care, have undergone rigorous testing and have been shown to be effective in treating COVID-19. While complementary and alternative medicinal plants may have potential benefits, they should be used in combination with, and not as a replacement for, conventional treatments.

4. Results & discussion

To find the mutations, our work consists of an alignment of protein sequences using bioinformatics tools like bioedit. In order to provide various fundamental functions like editing, aligning, manipulating, and analysing protein and nucleic sequences, BioEdit is a biological sequence editor that works on Windows. Although it lacks the capacity of more robust sequence analysis applications. BioEdit tool provides a number of quick and simple functions for annotating, editing, and manipulating sequences. Genomic sequence alignment is a way of arranging the protein sequences or DNA sequences to figure out similar regions which may be a reason of evolutionary relationships between the genomic sequences (Hall, 2004). Alignment of sequences has been performed to find the point mutations. Figure 2 shows the alignment of replicase protein sequences, here multiple replicase proteins have been considered which is found common for SARS-CoV2 different variants. The tool will align the sequences and save the aligned sequences in FASTA format. A phylogenetic tree is a branching diagram that represents the evolutionary relationships among a set of organisms or sequences, based on the similarities and differences in their genetic or protein sequences. In the case of SARS-CoV-2, a phylogenetic tree can be constructed using the amino acid sequences of the virus. The tree can help to visualize the evolutionary history of SARS-CoV-2, and can be used to identify the origin of the virus, its transmission patterns, and the emergence of new variants. The tree is typically constructed using bioinformatics software that can align the amino acid sequences, calculate

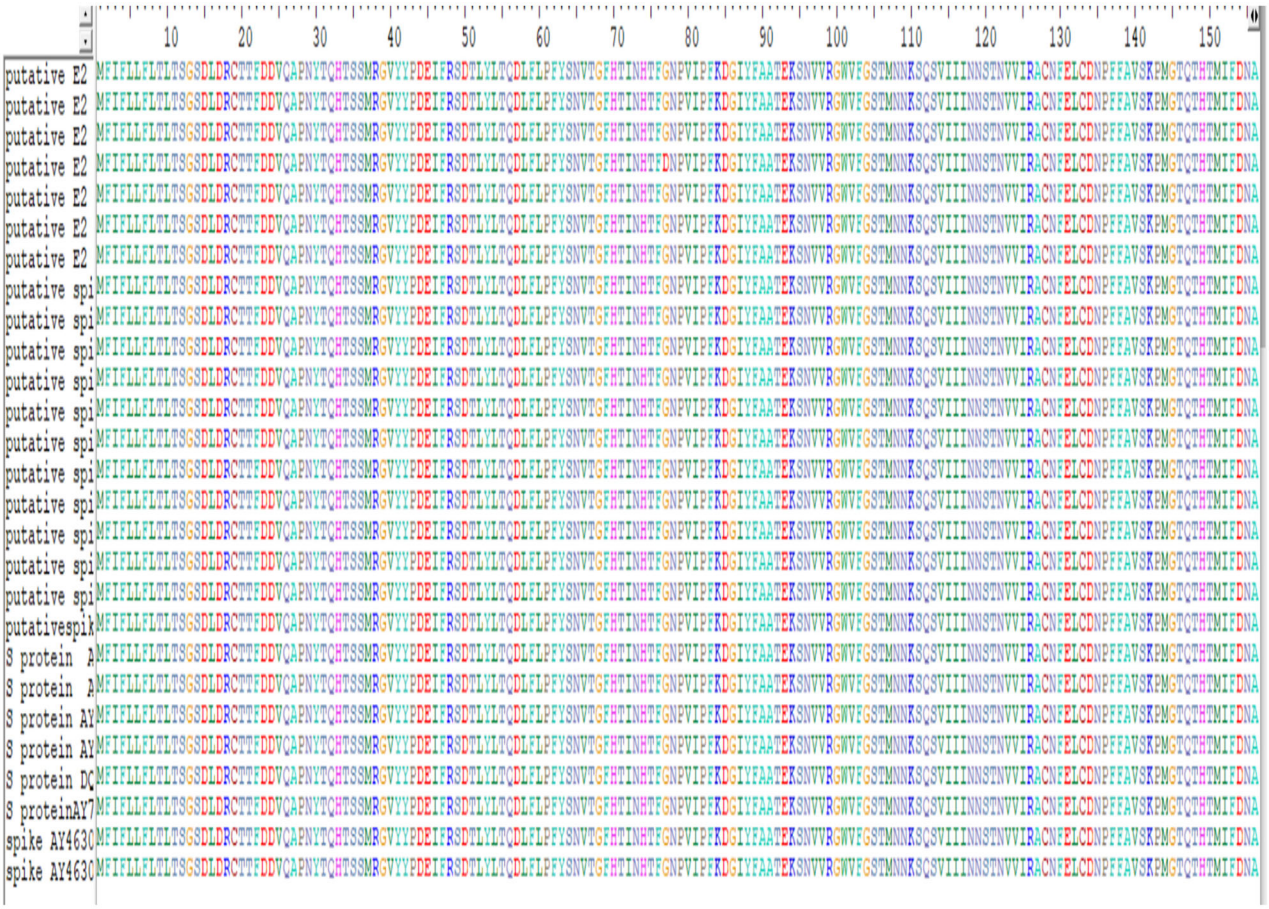

Figure 4. Alignment of putative spike GlycoProteins.

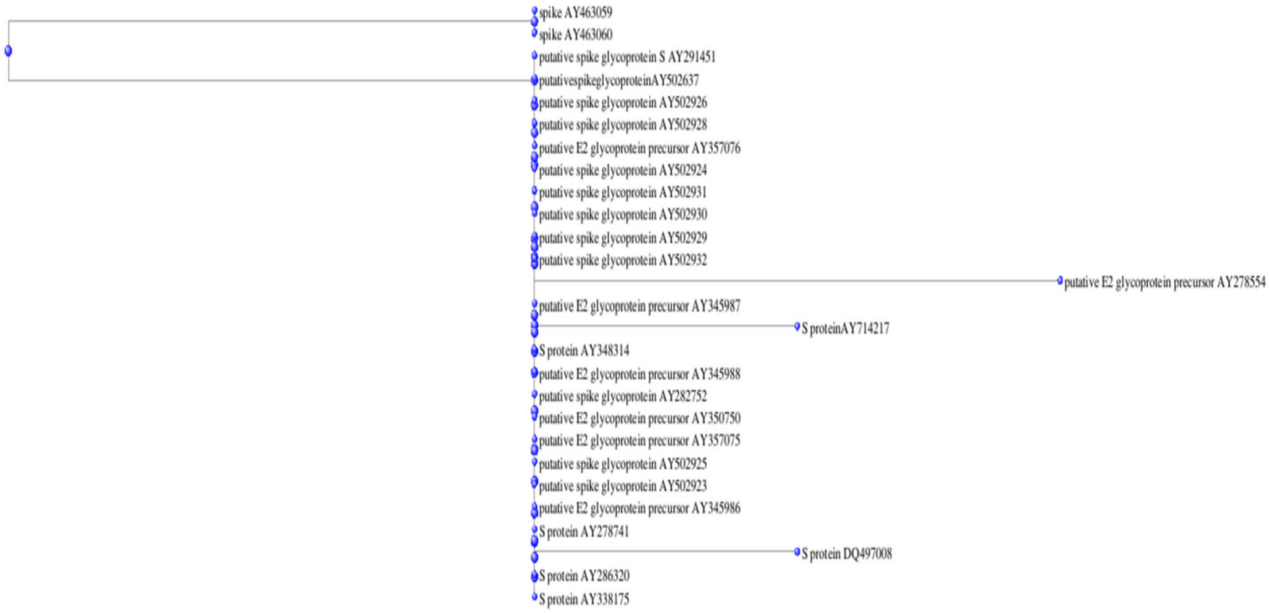

Figure 5. Phylogenetic tree of putative spike glycoproteins.

the genetic distances between them, and infer the branching patterns. One common software used to construct phylogenetic trees is MEGA (Molecular Evolutionary Genetics Analysis), which can handle large datasets and provide various options for phylogenetic analysis, including maximum likelihood, neighbor-joining, and Bayesian inference. Other software tools used for phylogenetic analysis include RAxML, PhyML, and BEAST. The resulting tree can be visualized using software such as FigTree or iTOL (Interactive Tree of Life), which allows for further customization and annotation of the tree. The phylogenetic tree can provide valuable insights into the evolution and diversity of SARS-CoV-2, and can inform public health measures and vaccine development strategies. Figure 3 shows the phylogenetic or

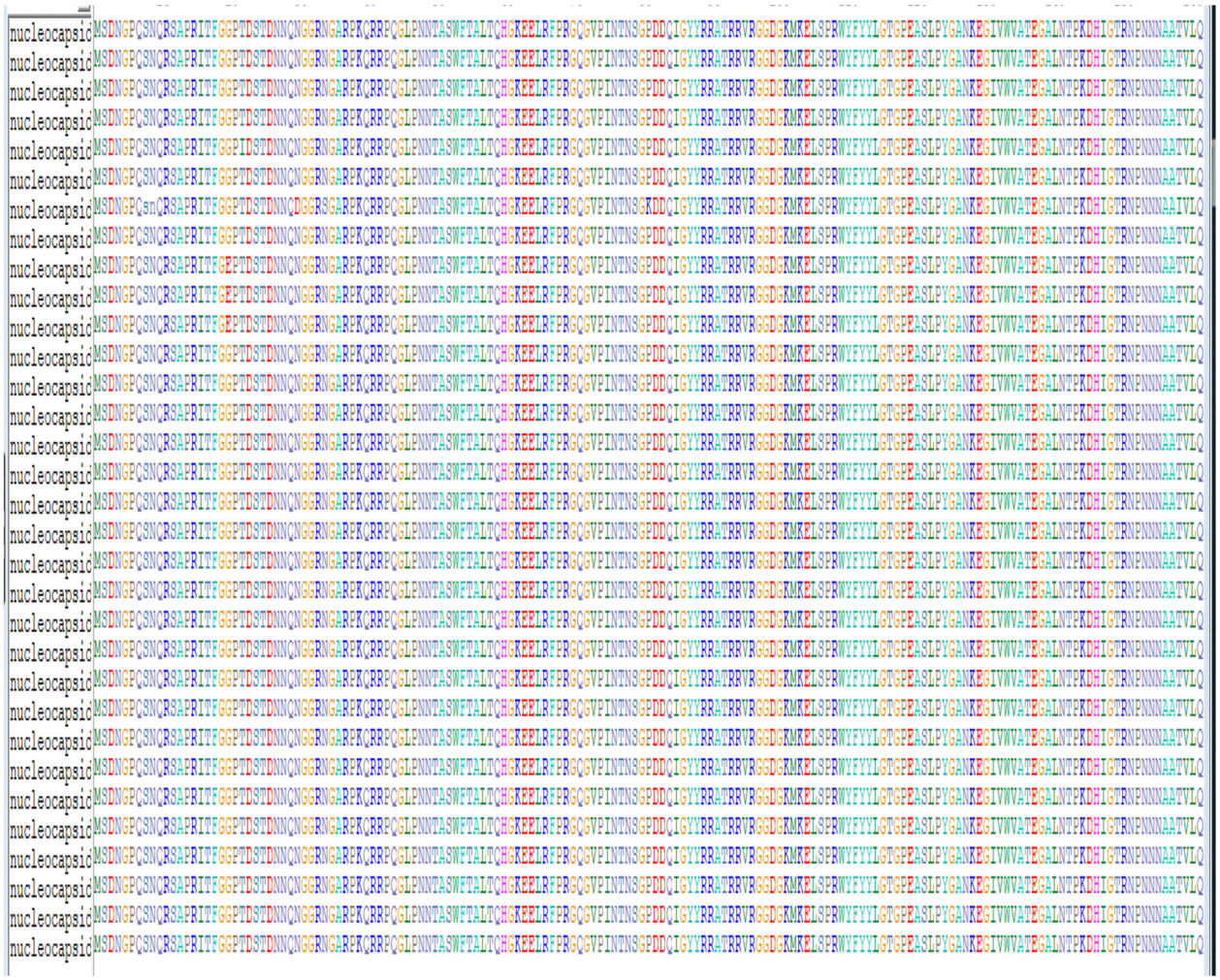

Figure 6. Alignment of nucleocapsid proteins.

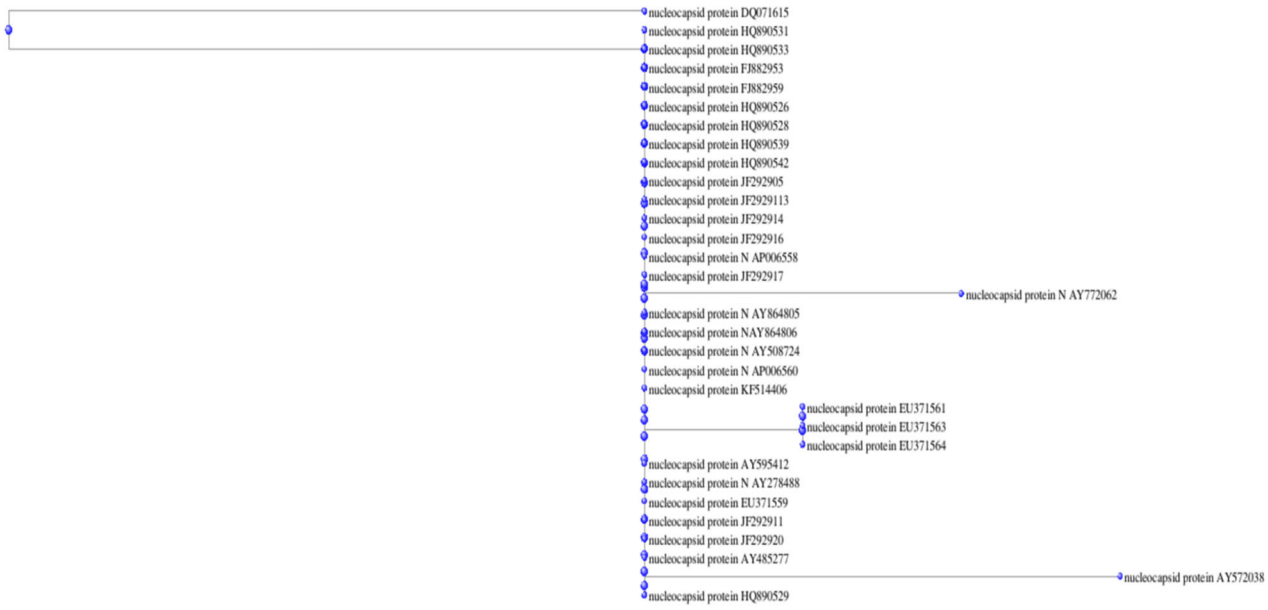

Figure 7. Phylogenetic tree of nucleocapsid proteins.

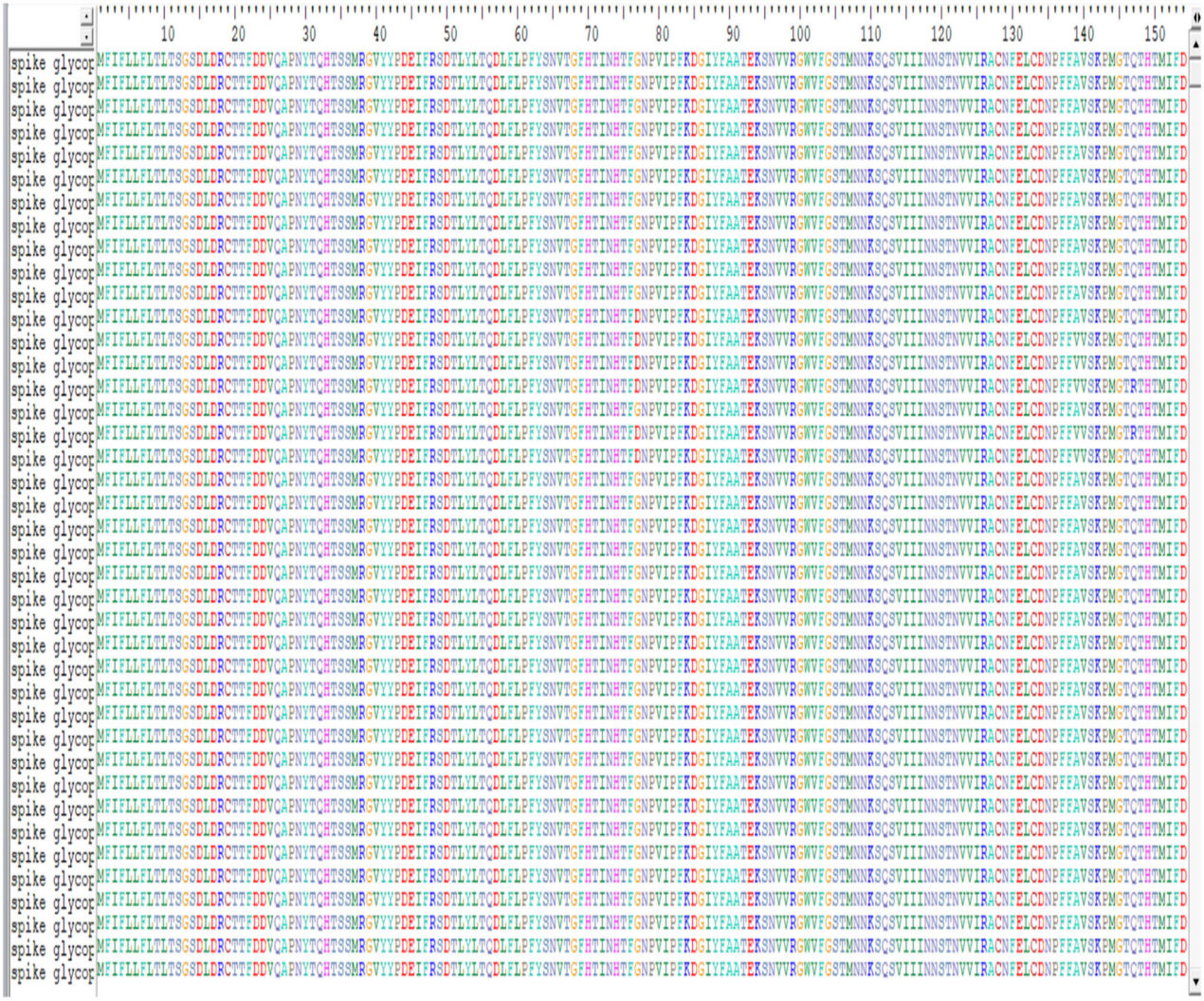

Figure 8. Alignment of spike proteins.

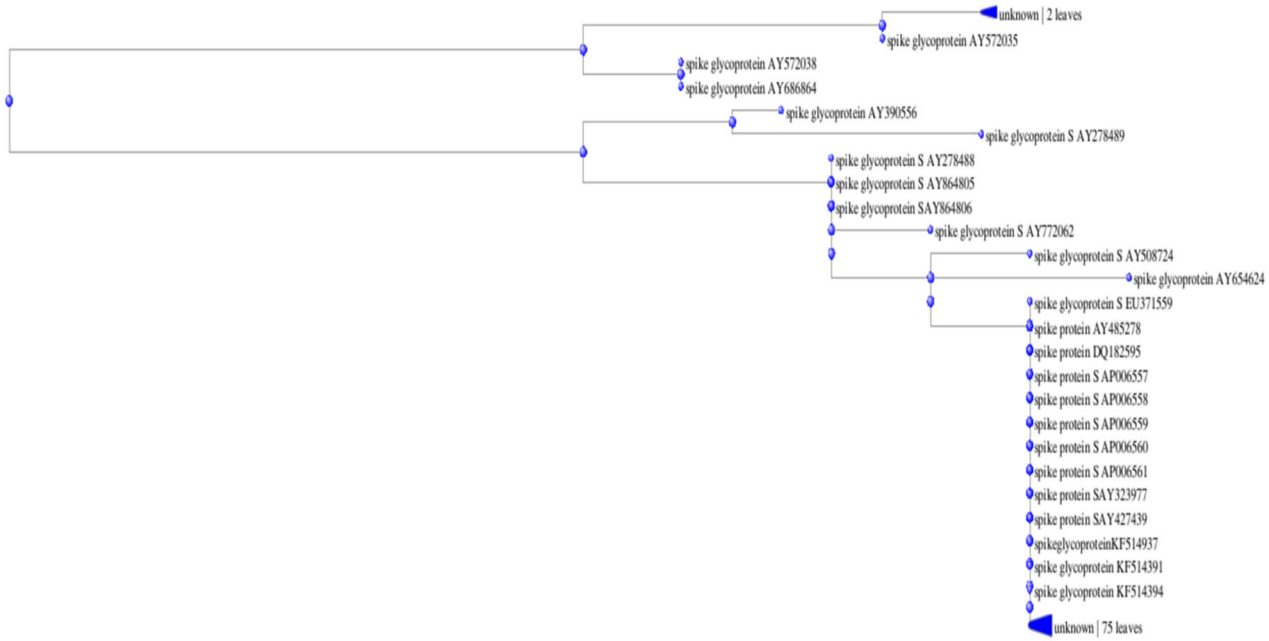

Figure 9. Phylogenetic tree of spike proteins.

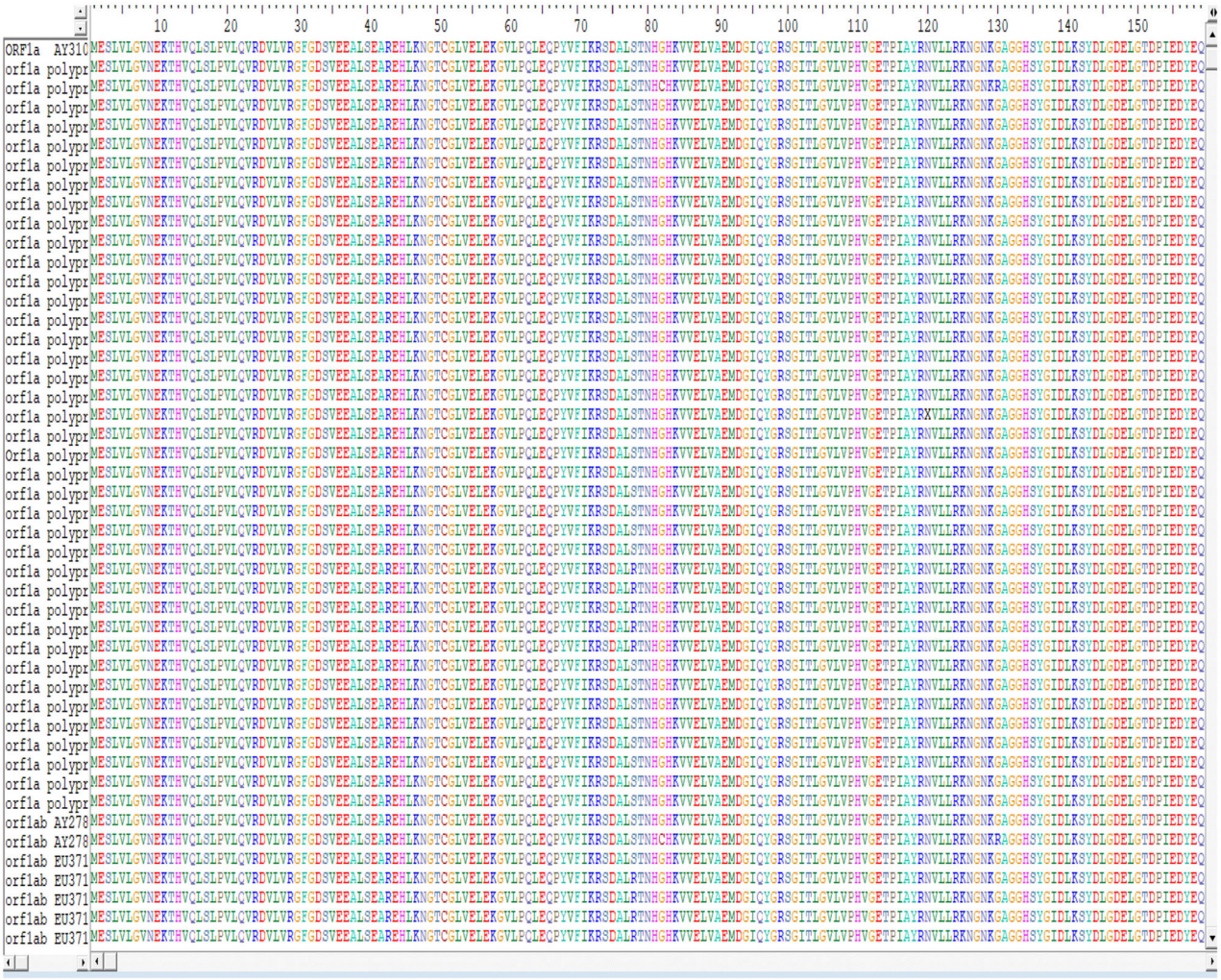

Figure 10. Alignment of ORF1a proteins.

evolution tree of replicase proteins which estimates the relationships among the sequences. This estimation can help in prescribing vaccines against them. This might give birth to new treatment options and also standing the progression of the virus.

The study of the link between biological lineages that have a common ancestor is known as phylogeny. To infer phylogeny, the differences between aligned sequences of genomes and proteins are measured and presented in the form of a tree, with modern species, intermediates, and common ancestors occupying the terminal nodes, internal nodes, and root, respectively. The tree's topology, branch length, shape, and root position are distinct features (Gorbalenya & Lauber, 2017).

Figure 4 shows the alignment of Spike GlycoProtein sequences, here multiple GlycoProtein proteins have been considered which is found common for SARS-CoV2 different variants. The tool will align the sequences and save the aligned sequences in FASTA format. Figure 5 shows the phylogenetic or evolution tree of Spike GlycoProteins.

Figure 6 shows the alignment of Nucleocapsid protein sequences, here multiple Nucleocapsid proteins have been considered which is found common for SARS-CoV2 different variants. The tool will align the sequences and save the aligned sequences in FASTA format. Figure 7 shows the phylogenetic or evolution tree of Nucleocapsid proteins.

Figure 8 shows the alignment of Spike protein sequences, here multiple Spike proteins have been considered which is found common for SARS-CoV2 different variants. The tool will align the sequences and save the aligned sequences in FASTA format and Figure 9 shows the phylogenetic or evolution tree of Spike proteins.

Figure 10 shows the alignment of ORF1a protein sequences, here multiple ORF1a proteins have been considered which is found common for SARS-CoV2 different variants. The tool will align the sequences and save the aligned sequences in FASTA format. and Figure 11 shows the phylogenetic or evolution tree of ORF1a proteins.

Now, after alignment and generation of evolution tree for the protein sequences, the next step Seq-2-Seq LSTM based encoder-decoder model is proposed in work. To predict the amino acid sequences and the future mutations deep learning approach is

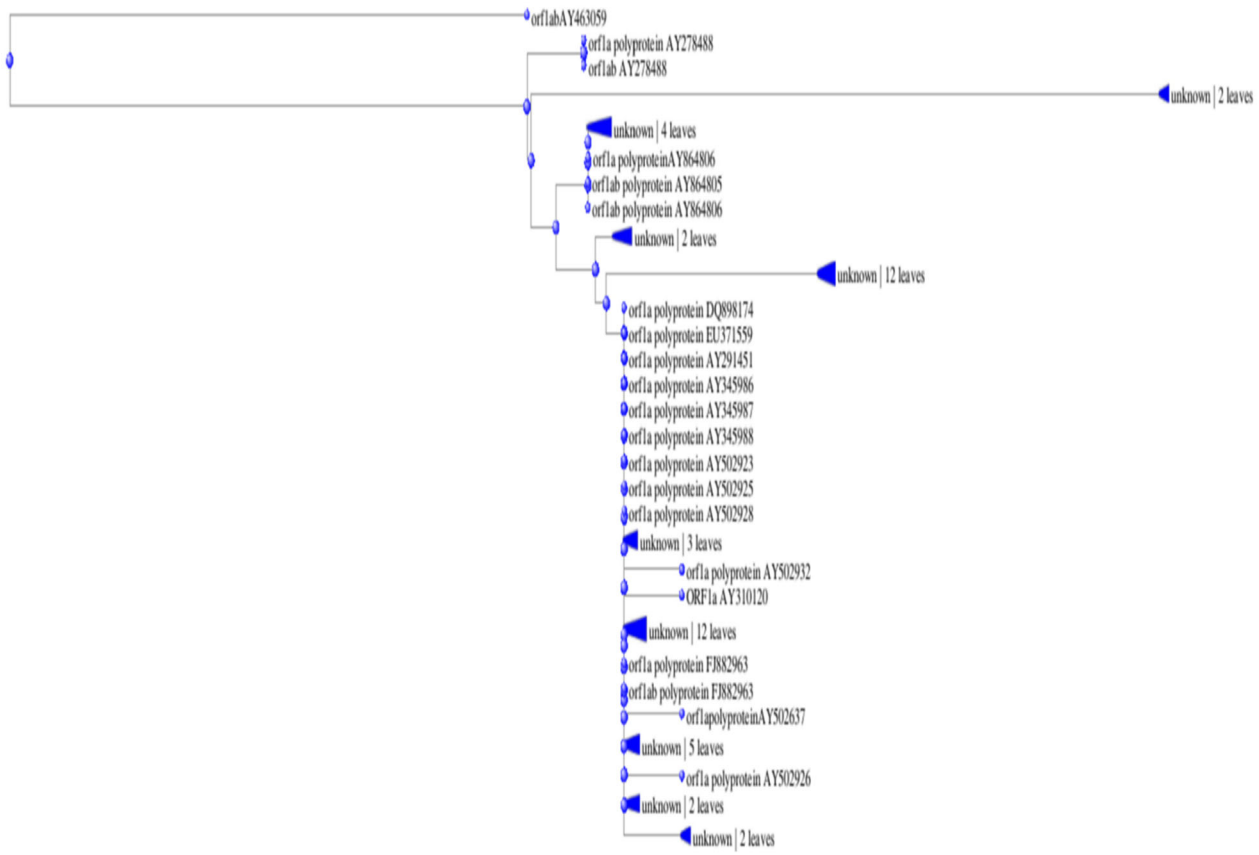

Figure 11. Phylogenetic tree of ORF1a proteins.

Table 4. Training of various protein sequences [sample].

	Spike Protein		Nucleocapsid		Replicase		Polyprotein		Putative		ORF1a	
Epoch	Time taken to perform each step (in ms)	Loss	Time taken to perform each step (in ms)	Loss	Time taken to perform each step (in ms)	Loss	Time taken to perform each step (in ms)	Loss	Time taken to perform each step (in ms)	Loss	Time taken to perform each step (in ms)	Loss
1	565	2.9984	650	3.0332	638	2.9703	634	2.9689	537	3.0166	662	2.9675
2	572	2.9791	648	2.9901	637	2.9579	640	2.9567	537	2.9880	671	2.9540
3	573	2.9118	640	2.9819	637	2.9492	636	2.9510	547	2.9828	741	2.8866
4	569	2.5597	639	2.9774	640	2.9232	638	2.9296	548	2.9797	657	2.5805
5	577	2.9463	641	2.9679	642	2.8284	641	2.8502	548	2.9784	654	2.6520
6	579	2.6987	641	2.9430	639	2.2152	644	2.2895	546	2.9758	645	2.9682
7	577	1.8095	642	2.8846	642	2.7760	642	2.9631	548	2.9619	652	2.9458
8	578	0.9822	646	2.8846	615	2.9826	640	2.9737	548	2.9265	663	2.9371
9	579	1.5924	641	2.7742	638	2.9430	640	2.9426	547	2.8925	650	2.9278
10	577	1.4703	642	2.6512	638	2.9255	643	2.9294	544	2.8475	652	2.6957
11	576	1.6648	650	2.6441	655	2.9094	642	2.9242	544	2.6904	645	2.6570
12	579	0.3274	639	2.6511	668	2.8832	642	2.9330	548	2.3366	650	2.5432
13	588	0.1482	639	2.6432	642	2.9188	645	2.8974	545	1.6613	657	2.4531
14	569	0.0875	641	2.6452	638	2.8148	656	2.8749	547	2.8181	662	2.5641
15	591	0.0721	639	2.6532	639	2.2222	663	2.8581	545	2.9816	659	2.6754
16	590	0.0499	638	2.6511	645	1.3474	669	2.8508	544	2.8914	651	2.5421
17	587	0.0442	638	2.6511	638	0.7237	657	2.4815	545	2.6995	655	2.7654
18	587	0.0394	638	2.6452	629	3.0311	659	1.5647	548	2.1195	653	2.4315
19	589	0.0419	639	2.6511	633	2.9346	653	0.7270	547	1.4188	678	2.2314
20	592	0.0322	639	2.6511	631	2.8619	643	0.3370	549	1.0359	654	2.1800

used. Below is Table 4. This table shows that the various protein sequences are being trained to predict the future sequence mutation, which consists of the time taken to perform each step and the associated loss. The model is trained for 50 epochs, with batch size 10 for optimization. The model consists of adam optimizer. The learning rate considered is 0.001 with 100 hidden neurons. The dropout value considered is 0.5. The proposed model is trained in Colab. The metric considered is accuracy for model performance

Below we can see the similarity between various amino acid sequences or protein sequences. The analysis should show that the sequences were more

Table 5. Average similarity percentage of amino acid sequences (pairwise).

Proteins	Similarity between Amino Acid Sequences of various SARS-Coronavirus (%)
Nucleocapsid protein	99.9
ORF1a	99.8
Putative Spike Glycoprotein	99.7
Replicase	99.8
Spike Glycoprotein	99.9

similar to each other. Very few mutations were found in the sequences. This result is based on the dataset which is used. Table 5 shows the average pairwise similarity percentage between amino acid sequences of SARS-Cov2.

Finally, the predicted amino acid sequences are approximately 98% similar to the trained amino acid sequence, and the mutations observed were negligible due to the high similarity between the amino acid sequences. The accuracy metrics basically determine the correct predictions that a trained deep-learning model achieves.

5. Conclusion

In summary, mutations in SARS-CoV-2 can have significant impacts on pathogenicity, diagnostics, therapeutics, and vaccines. It is important for researchers and public health officials to monitor the evolution of the virus and its mutations to ensure that diagnostic tests, treatments, and vaccines remain effective. Deep learning algorithms play a significant role in bioinformatics. Various deep learning algorithms can be used to do tasks such as sequence categorization and prediction in a short amount of time. Our research focused on predicting mutations and computing similarities between protein sequences. The proposed LSTM model obtained an average prediction accuracy of 98% for all protein sequences included in the study, that is, spike, replicase, putative, ORF1a, Nucleocapsid, and PolyProtein. This prediction is beneficial in developing drugs for specific altered protein sequences. Finally, this study has shown that SARS-CoV2 sequence prediction is possible in the future. Bioinformatics tools and a deep learning-based technique were used to examine and visualize amino acid similarities, mutation prediction, and phylogenetic analysis. Despite various research still, there are various other challenges also, like considering the global information of proteins and finding the changes in predicting mutation.

Authors' contributions

All the authors have equal contributions to completing the manuscript.

Availability of data and materials

All the data available with authors, we will supply as demand comes from the reviewer.

Disclosure statement

There is no conflict of Interest between the author.

ORCID

Sachi Nandan Mohanty http://orcid.org/0000-0002-4939-0797

References

Carvalho, P. C., Fischer, J. S., & Chen, E. I. (2009). DomProtein explorer: A tool for exploring domain-domain interactions in protein structures. *Bioinformatics*, *25*(9), 1235–1236.

Centers for Disease Control and Prevention (2021). Emerging SARS-CoV-2 variants. https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html.

Chen, J., Gao, K., Wang, R., & Wei, G.-W. (2021). Prediction and mitigation of mutation threats to COVID-19 vaccines and antibody therapies. *Chemical Science*, *12*(20), 6929–6948. doi:10.1039/d1sc01203g

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. doi:10.1093/nar/gkh340

Gorbalenya, A. E., & Lauber, C. (2017). Phylogeny of viruses reference module in biomedical sciences.

Hall, T. (1999). BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95–98.

Hall, T. (2004). "BioEdit version 7.0. 0." Distributed by the author, website: www. mbio. ncsu. edu/BioEdit/bioedit. Html.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., … Bhattacharya, T. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. https://www.biorxiv.org/content/10.1101/2020.04.29.069054v2.

Koumakis, L. (2020). Deep learning models in genomics. *Computational and Structural Biotechnology Journal*, *18*, 1466–1473. doi:10.1016/j.csbj.2020.06.017

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. doi:10.1093/molbev/msw054

Lauring, A. S., & Hodcroft, E. B. (2021). Genetic variants of SARS-CoV-2-what do they mean? *JAMA*, *325*(6), 529–531. doi:10.1001/jama.2020.27124

Lopez-Rincon, A., Tonda, A., Mendoza-Maldonado, L., Mulders, D. G. J. C., Molenkamp, R., Perez-Romero, C. A., … Kraneveld, A. D. (2021). Classi_cation and speci_c primer design for accurate detection of SARS-CoV-2 using deep learning. *Sci. Rep., Vol*, *11*(1), 1–11.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., … Bi, Y. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet*, *395*(10224), 565–574.

Mohamed, T., Sayed, S., Salah, A., Houssein, E. H. (2021). Long short-term memory neural networks for RNA viruses mutations prediction. *Mathematical Problems in Engineering*, Article ID 9980347, 9. doi:10.1155/2021/9980347

National Center for Biotechnology Information (NCBI) Bethesda (MD). (1988). National Library of Medicine (US), National Center for Biotechnology Information; https://www.ncbi.nlm.nih.gov/. Accessed 30 January 2022.

Nawaz, M. S., Fournier-Viger, P., Shojaee, A., & Fujita, H. (2021). Using artificial intelligence techniques for COVID-19 genome analysis. *Applied Intelligence (Dordrecht, Netherlands)*, *51*(5), 3086–3103. doi:10.1007/s10489-021-02193-w

Nguyen, T. T., Pathirana, P. N., Nguyen, T., Nguyen, Q. V. H., Bhatti, A., Nguyen, D. C., … Abdelrazek, M. (2021). Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus). *Scientific Reports*, *11*(1), 1–16.

Pathan, R. K., Biswas, M., & Khandaker, M. U. (2020). Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model. *Chaos, Solitons, and Fractals*, *138*, 110018. doi:10.1016/j.chaos.2020.110018

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., … Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology*, *5*(11), 1403–1407. doi:10.1038/s41564-020-0770-5

Sah, S., Dr.Surendiran, B., Dr.Dhanalakshmi, R., & Kamerkar, A. (2021). Classification and alignment of SARS-CoV2 sequences using machine learning approach. *International Journal of Advanced Research in Management, Architecture, Technology and Engineering*, *7*, 34–44.

Sah, S., Surendiran, B., & Dhanalakshmi, R. (2023). Genomic sequence similarity of SARS-CoV2 nucleotide sequences using biopython: Key for finding cure and vaccines. In *Application of deep learning methods in healthcare and medical science* (pp. 211–223). USA: Apple Academic Press.

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. doi:10.1038/nbt1486

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science: A Publication of the Protein Society*, *27*(1), 135–145. doi:10.1002/pro.3290

Smith, Y. (2019). Amino acids and protein sequences news. https://www.news-medical.net/life-sciences/Amino-Acids-and-Protein-Sequences.aspx. Accessed 26 Feb 2019

Stranger, B. E., & Dermitzakis, E. T. (2006). From DNA to RNA to disease and back: The 'central dogma' of regulatory disease variation Hum. *Genomics*, *2*(6), 383–390.

Taly, J. F., Magis, C., Bussotti, G., Chang, J. M., Di Tommaso, P., Erb, I., … Notredame, C. (2011). The coffee served blind: A new view on the multiple sequence alignment problem. *PLoS One*. *6*(12), e28817. doi:10.1371/journal.pone.0028817

Tomita, N., Mori, H., & Mochizuki, T. (2015). An efficient way of selecting multiple sequences for BioEdit. *Bioscience, Biotechnology, and Biochemistry*, *79*(12), 2013–2015.

Wang, R., Chen, J., Gao, K., & Wei, G.-W. (2021). Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries. *Genomics*, *113*(4), 2158–2170. doi:10.1016/j.ygeno.2021.05.006

Whata, A., & Chimedza, C. (2021). Deep learning for SARS COV-2 genome sequences. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 59597–59611. doi:10.1109/ACCESS.2021.3073728

World Health Organization (2021). Tracking SARS-CoV-2 variants. https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/.

Xu, J., Guo, H. C., Wei, Y. Q., Shu, L., Wang, J., Li, J. S., … Sun, S. Q. (2015). Phylogenetic analysis of canine parvovirus isolates from Sichuan and Gansu provinces of China in 2011. *Transboundary and Emerging Diseases*, *62*, 91–95.

Yan, S., & Wu, G. (2021). Neural network to predict probabilistically possible mutations in hemagglutinins from Eurasia H1 influenza A virus. In *2nd International Conference on Computer Vision, Image, and Deep Learning*, vol. 11911, pp. 283–289. SPIE.

Zhou, B., Zhou, H., Zhang, X., Xu, X., Chai, Y., Zheng, Z., … Zhou, Z. (2023). TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution. *Computers in Biology and Medicine*, *152*, 12–21.

Annexure 1

Table A1. Detailed description of protein sequence data.

Protein Name	Aligned Protein Sequences with accession number	Each Sequence Length after Alignment
Nucleocapsid	**>nucleocapsid protein AY485277** **>nucleocapsid protein HQ890531** **>nucleocapsid protein HQ890533** **>nucleocapsid protein KF514406** **>nucleocapsid protein AY572038** **>nucleocapsid protein AY595412** **>nucleocapsid protein DQ071615** **>nucleocapsid protein EU371559** **>nucleocapsid protein EU371561** **>nucleocapsid protein EU371563** **>nucleocapsid protein EU371564** **>nucleocapsid protein FJ882953** **>nucleocapsid protein FJ882959** **>nucleocapsid protein HQ890526** **>nucleocapsid protein HQ890528** **>nucleocapsid protein HQ890529** **>nucleocapsid protein HQ890539** **>nucleocapsid protein HQ890542** **>nucleocapsid protein JF292905** **>nucleocapsid protein JF292911** **>nucleocapsid protein JF2929113** **>nucleocapsid protein JF292914** **>nucleocapsid protein JF292916** **>nucleocapsid protein JF292917** **>nucleocapsid protein JF292920** **>nucleocapsid protein N AY278488** **>nucleocapsid protein N AP006558** **>nucleocapsid protein N AP006560** **>nucleocapsid protein N AY508724** **>nucleocapsid protein N AY772062** **>nucleocapsid protein N AY864805** **>nucleocapsid protein NAY864806**	426
Putative spike gylcoprotein	**>putative E2 glycoprotein precursor AY345986** **>putative E2 glycoprotein precursor AY345987** **>putative E2 glycoprotein precursor AY345988** **>putative E2 glycoprotein precursor AY278554** **>putative E2 glycoprotein precursor AY350750** **>putative E2 glycoprotein precursor AY357075** **>putative E2 glycoprotein precursor AY357076** **>putative spike glycoprotein AY282752** **>putative spike glycoprotein AY502923** **>putative spike glycoprotein AY502924** **>putative spike glycoprotein AY502925** **>putative spike glycoprotein AY502926** **>putative spike glycoprotein AY502928** **>putative spike glycoprotein AY502929** **>putative spike glycoprotein AY502930** **>putative spike glycoprotein AY502931** **>putative spike glycoprotein AY502932** **>putative spike glycoprotein S AY291451** **>putativespikeglycoproteinAY502637** **>S protein AY278741** **>S protein AY286320** **>S protein AY338175** **>S protein AY348314** **>S protein DQ497008** **>S proteinAY714217** **>spike AY463059** **>spike AY463060**	1275
Replicase	**>replicase 1AB AY274119** **>replicase 1ab AY654624** **>replicase 1AB AY686863** **>replicase 1AB AY686864** **>replicase 1AB polyprotein FJ959407** **>replicase p1AB AY572034** **>replicase p1AB AY572035** **>replicase polyprotein 1a KF514403** **>replicase polyprotein 1a KF514404** **>replicase polyprotein 1a KF514406** **>replicase polyprotein 1a KF514409** **>replicase polyprotein 1a KF514411** **>replicase polyprotein 1a KF514415** **>replicase polyprotein 1a KF514418**	4448

(*continued*)

Table A1. Continued.

Protein Name	Aligned Protein Sequences with accession number	Each Sequence Length after Alignment
	>replicase polyprotein 1a KF514420	
	>replicase polyprotein 1a KF514422	
	>replicase polyprotein 1a KF514423	
	>replicase polyprotein 1a KF514388	
	>replicase polyprotein 1a KF514391	
	>replicase polyprotein 1a KF514394	
	>replicase polyprotein 1a KF514398	
	>replicase polyprotein 1a KF514399	
	>replicase polyprotein 1a KF514400	
	>replicase polyprotein 1ab KF514399	
	>replicase polyprotein 1ab KF514403	
	>replicase polyprotein 1ab KF514404	
	>replicase polyprotein 1ab KF514406	
	>replicase polyprotein 1ab KF514409	
	>replicase polyprotein 1ab KF514411	
	>replicase polyprotein 1ab KF514415	
	>replicase polyprotein 1ab KF514418	
	>replicase polyprotein 1ab KF514420	
	>replicase polyprotein 1ab KF514388	
	>replicase polyprotein 1ab KF514389	
	>replicase polyprotein 1ab KF514391	
	>replicase polyprotein 1ab KF514394	
	>replicase polyprotein 1ab KF514398	
	>replicase polyprotein 1ab KF514400	
	>replicase polyprotein 1ab KF514422	
	>replicase polyprotein 1ab KF514423	
	>replicase polyprotein 1aKF514389	
	>replicase1ABNC004718	
	>replicasepolyprotein1abKF514937	
	>replicasepolyprotein1aKF514937	
ORF1a	**>ORF1a AY310120**	68
	>orf1a polyprotein AY278488	
	>orf1a polyprotein AY278489	
	>orf1a polyprotein AY286320	
	>orf1a polyprotein AY291451	
	>orf1a polyprotein AY345986	
	>orf1a polyprotein AY345987	
	>orf1a polyprotein AY345988	
	>orf1a polyprotein AY485277	
	>orf1a polyprotein AY485278	
	>orf1a polyprotein AY502923	
	>orf1a polyprotein AY502924	
	>orf1a polyprotein AY502925	
	>orf1a polyprotein AY502926	
	>orf1a polyprotein AY502928	
	>orf1a polyprotein AY502929	
	>orf1a polyprotein AY502930	
	>orf1a polyprotein AY502931	
	>orf1a polyprotein AY502932	
	>orf1a polyprotein AY508724	
	>orf1a polyprotein AY772062	
	>orf1a polyprotein AY864805	
	>orf1a polyprotein DQ182595	
	>orf1a polyprotein DQ898174	
	>orf1a polyprotein EU371559	
	>orf1a polyprotein EU371560	
	>orf1a polyprotein EU371561	
	>orf1a polyprotein EU371562	
	>orf1a polyprotein EU371563	
	>orf1a polyprotein EU371564	
	>orf1a polyprotein FJ882943	
	>orf1a polyprotein FJ882953	
	>orf1a polyprotein FJ882958	
	>orf1a polyprotein FJ882959	
	>orf1a polyprotein FJ882961	
	>orf1a polyprotein FJ882962	
	>orf1a polyprotein FJ882963	
	>orf1a polyproteinAY864806	
	>orf1ab AY278488	
	>orf1ab AY278489	
	>orf1ab EU371559	
	>orf1ab EU371560	
	>orf1ab EU371561	
	>orf1ab EU371562	
	>orf1ab EU371563	
	>orf1ab EU371564	
	>orf1ab polyprotein AY286320	

(*continued*)

Table A1. Continued.

Protein Name	Aligned Protein Sequences with accession number	Each Sequence Length after Alignment
	>orf1ab polyprotein AY485277	
	>orf1ab polyprotein AY485278	
	>orf1ab polyprotein AY772062	
	>orf1ab polyprotein AY864805	
	>orf1ab polyprotein AY864806	
	>orf1ab polyprotein DQ182595	
	>orf1ab polyprotein DQ898174	
	>orf1ab polyprotein FJ882943	
	>orf1ab polyprotein FJ882953	
	>orf1ab polyprotein FJ882958	
	>orf1ab polyprotein FJ882959	
	>orf1ab polyprotein FJ882961	
	>orf1ab polyprotein FJ882962	
	>orf1ab polyprotein FJ882963	
	>orf1abAY463059	
	>orf1apolyproteinAY502637	
Spike	**>spike glycoprotein AY274119**	103
	>spike glycoprotein AY310120	
	>spike glycoprotein KF514403	
	>spike glycoprotein KF514404	
	>spike glycoprotein KF514406	
	>spike glycoprotein KF51440620	
	>spike glycoprotein KF514409	
	>spike glycoprotein KF514411	
	>spike glycoprotein KF514415	
	>spike glycoprotein KF514418	
	>spike glycoprotein KF514422	
	>spike glycoprotein AY390556	
	>spike glycoprotein AY572034	
	>spike glycoprotein AY572035	
	>spike glycoprotein AY572038	
	>spike glycoprotein AY654624	
	>spike glycoprotein AY686864	
	>spike glycoprotein AY686863	
	>spike glycoprotein DQ898174	
	>spike glycoprotein KF514388	
	>spike glycoprotein KF514389	
	>spike glycoprotein KF514391	
	>spike glycoprotein KF514394	
	>spike glycoprotein KF514398	
	>spike glycoprotein KF514399	
	>spike glycoprotein KF514400	
	>spike glycoprotein KF514423	
	>spike glycoprotein precursor HQ890530	
	>spike glycoprotein precursor HQ890531	
	>spike glycoprotein precursor HQ890532	
	>spike glycoprotein precursor HQ890533	
	>spike glycoprotein precursor HQ890534	
	>spike glycoprotein precursor HQ890535	
	>spike glycoprotein precursor HQ890536	
	>spike glycoprotein precursor HQ890537	
	>spike glycoprotein precursor HQ890538	
	>spike glycoprotein precursor FJ882943	
	>spike glycoprotein precursor FJ882953	
	>spike glycoprotein precursor FJ882958	
	>spike glycoprotein precursor FJ882959	
	>spike glycoprotein precursor FJ882961	
	>spike glycoprotein precursor FJ882962	
	>spike glycoprotein precursor FJ882963	
	>spike glycoprotein precursor HQ890526	
	>spike glycoprotein precursor HQ890528	
	>spike glycoprotein precursor HQ890529	
	>spike glycoprotein precursor HQ890539	
	>spike glycoprotein precursor HQ890540	
	>spike glycoprotein precursor HQ890541	
	>spike glycoprotein precursor HQ890542	
	>spike glycoprotein precursor HQ890543	
	>spike glycoprotein precursor HQ890545	
	>spike glycoprotein precursor HQ890546	
	>spike glycoprotein precursor JF292903	
	>spike glycoprotein precursor JF292904	
	>spike glycoprotein precursor JF292905	
	>spike glycoprotein precursor JF292906	
	>spike glycoprotein precursor JF292907	
	>spike glycoprotein precursor JF292908	
	>spike glycoprotein precursor JF292910	
	>spike glycoprotein precursor JF292911	

(*continued*)

Table A1. Continued.

Protein Name	Aligned Protein Sequences with accession number	Each Sequence Length after Alignment
	>spike glycoprotein precursor JF292913	
	>spike glycoprotein precursor JF292914	
	>spike glycoprotein precursor JF292915	
	>spike glycoprotein precursor JF292916	
	>spike glycoprotein precursor JF292917	
	>spike glycoprotein precursor JF292918	
	>spike glycoprotein precursor JF292919	
	>spike glycoprotein precursor JF292920	
	>spike glycoprotein precursor JF292921	
	>spike glycoprotein precursor JF292922	
	>spike glycoprotein precursorHQ890527	
	>spike glycoprotein precursorJF292902	
	>spike glycoprotein S AY278488	
	>spike glycoprotein S AY278489	
	>spike glycoprotein S AY508724	
	>spike glycoprotein S AY772062	
	>spike glycoprotein S AY864805	
	>spike glycoprotein S EU371559	
	>spike glycoprotein S EU371560	
	>spike glycoprotein S EU371561	
	>spike glycoprotein S EU371562	
	>spike glycoprotein S EU371563	
	>spike glycoprotein S EU371564	
	>spike glycoprotein SAY864806	
	>spike protein AY485277	
	>spike protein AY485278	
	>spike protein DQ182595	
	>spike protein S AY291315	
	>spike protein S AP006557	
	>spike protein S AP006558	
	>spike protein S AP006559	
	>spike protein S AP006560	
	>spike protein S AP006561	
	>spike protein SAY323977	
	>spike protein SAY427439	
	>spikeglycoproteinKF514937	
	>spikeglycoproteinNC004718	
	>spikeproteinAB257344	
	>spikeglycoproteinprecursorJF292912	