Deep Recurrent Neural Networks for the Generation of Synthetic Coronavirus Spike Protein Sequences

Lisa C. Crossman[1,2(✉)]

[1] SequenceAnalysis.co.uk, NRP Innovation Centre, Norwich Research Park, Norwich, UK
l.crossman@uea.ac.uk
[2] School of Biological Sciences, University of East Anglia, Norwich, Norfolk, UK

Abstract. With the advent of deep learning techniques for text generation, comes the possibility of generating fully simulated or synthetic genomes. For this study, the dataset of interest is that of coronaviruses. Coronaviridae are a family of positive-sense RNA viruses capable of infecting humans and animals. These viruses usually cause mild to moderate upper respiratory tract infection; however, they can also cause more severe symptoms, gastrointestinal and central nervous system diseases. The viruses are capable of flexibly adapting to new environments, hence health threats from coronavirus are constant and long-term. Immunogenic spike proteins are glycoproteins found on the surface of Coronaviridae particles that mediate entry to host cells. The aim of this study was to train deep learning neural networks to produce simulated spike protein sequences, which may be able to aid in knowledge and/or vaccine design by creating alternative possible spike sequences that could arise from zoonotic sources in future. Deep learning recurrent neural networks (RNN) were trained to provide computer-simulated coronavirus spike protein sequences in the style of previously known sequences and examine their characteristics. The deep generative model was created as a recurrent neural network employing text embedding and gated recurrent unit layers in TensorFlow Keras. Training used a dataset of alpha, beta, gamma, and delta coronavirus spike sequences. In a set of 100 simulated sequences, all 100 had most significant BLAST matches to Spike proteins in searches against NCBI non-redundant dataset (NR) and possessed the expected Pfam domain matches. Simulated sequences from the neural network may be able to guide us with future prospective targets for vaccine discovery in advance of a potential novel zoonosis.

Keywords: Coronavirus · Deep learning · Neural networks

1 Introduction

1.1 Coronaviridae

Coronaviridae are a family of large, enveloped single-stranded positive-sense RNA viruses encompassing alpha, beta, gamma, and delta coronavirus divisions as well as unclassified divisions in the sequence databases. The genome is packed inside a helical

D. Chicco et al. (Eds.): CIBB 2021, LNBI 13483, pp. 217–226, 2022.
https://doi.org/10.1007/978-3-031-20837-9_17

capsid and is further surrounded by an envelope. The spike protein forms large protrusions from the virus surface, giving the coronaviruses the appearance of wearing a 'crown' under electron microscopy. Coronaviruses can infect a wide range of different animals and usually cause mild to moderate upper-respiratory tract illnesses, however they can also cause severe respiratory infections as well as gastrointestinal and central nervous system diseases. Coronaviruses circulate among humans and animals such as bats, pigs, camels, and cats. Recent zoonoses include severe acute respiratory syndrome Coronavirus (SARS-CoV), which emerged in November 2002 and became effectively extinct by 2004 [1]. Another zoonosis, Middle East Respiratory Syndrome (MERS-CoV) was believed to be transmitted from an animal reservoir in camels in 2012 [2]. In veterinary terms, economically important CoV exist such as porcine epidemic diarrhoea coronavirus (PEDV) which lead to an extremely high fatality rate in piglets [3]. The coronavirus SARS-CoV-2 emerged from China in 2019 [4] and was declared a pandemic during the first quarter of 2020 with an extremely high requirement for a vaccine to be provided in a short timeframe. The Spike protein is a multifunctional viral protein found on the outside of the SARS-CoV-2 virus particle (Fig. 1).

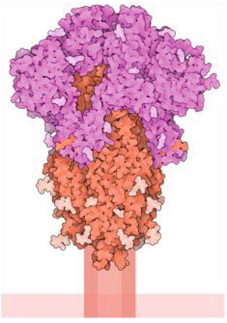

Fig. 1. Spike protein structure from SARS-CoV-2. The S1 fragment is shown in magenta, the S2 fragment is shown in red, with glycosylation as lighter hues. The receptor binding domain (RBD) is located at the top of the molecule, whilst the S1 and S2 fragments form part of a complex with a membrane-spanning segment. Image adapted from D. Goodsell and RCSB PDB [5]. (Color figure online)

Spike protein initially binds a host cell receptor though its S1 subunit and fuses viral and host membranes through its S2 subunit. In addition to mediating entry, the spike is a critical determinant of viral host range and a major inducer of host immune responses [6]. Due to the key role of the Spike (S) protein, it is the main target for antibody-mediated neutralization [7].

1.2 Recurrent Neural Networks

Deep learning is a subset of artificial intelligence employing neural networks. The recurrent neural network (RNN) is a type of neural network usually used for text encoding implementations, mainly through whole word encoding and the bag of words concept.

The recurrent neural network (RNN) is trained on a set of sequences using an optimization algorithm with estimations of gradient descent combined with backpropagation through time. The RNN has the potential to consider previously seen data such as the character or word that came before the current time step using units such as long short-term memory cells (LSTM) or gated recurrent units (GRU). The GRU is a variant of LSTM with a forget gate but having fewer parameters than LSTM as it lacks an output gate [8]. GRU performance is similar to LSTM but can be enhanced on some datasets.

In 2007, Hochreiter, Heusel and Obermayer proposed the use of LSTM for protein homology detection [9], commenting that LSTM is capable of automatically extracting local and global sequence statistics like hydrophobicity, polarity, volume and polarizability and combining them with a pattern. The results included extraction of feature dependencies that were not detected with common bioinformatic techniques. In this study, we investigate whether GRU is capable of learning these features in the context of generating synthetic sequences.

2 Methods

2.1 Recurrent Neural Network (RNN) Architecture

In creating the model described in this study, character encoding was used on the sequences in the training set. Alternative model architectures were considered which included either two layers of GRU, a single bidirectional layer of GRU, or two bidirectional layers of GRU. In addition, either a single dense layer was used as output, or two dense layers, with the first dense layer having half the number of RNN units (512). Model architecture changes also included swapping GRU for LSTM. However, the model showing the best results as judged by bioinformatic analysis of the output synthetic sequences was composed of a single embedding layer and a gated recurrent unit (GRU) with 1024 RNN units followed by a dense linear layer (Fig. 2).

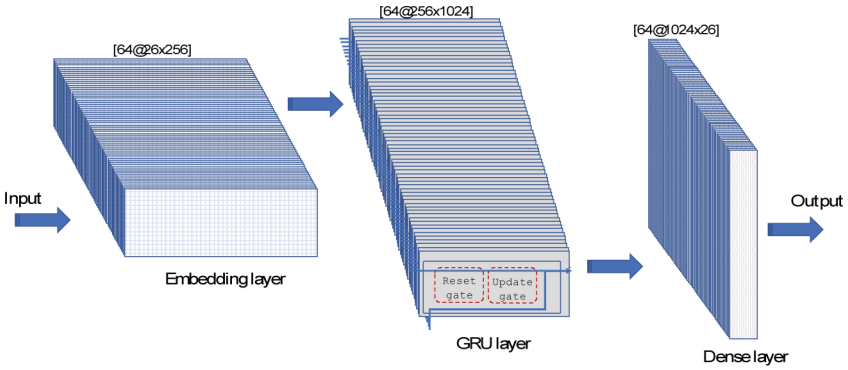

Fig. 2. Simple recurrent neural network trained for the production of synthetic protein sequences of Spike protein. The vocabulary size was 26, including each of the protein single letter codes, plus newline and space characters, the embedding dimension was 256, with a RNN units parameter of 1024. During prediction, after loading the trained model, the model is reset to a batch size of one and prediction is carried out one character at a time starting with the seed text.

The model was trained in Tensorflow 2.1.0 with Keras using an Adam optimizer with AMSgrad option and an adaptive learning rate over 15 epochs, where losses fell gradually from an initial 3.259 to 0.266. Learning was terminated after the losses had fallen between 0.2–0.3.

2.2 Coronavirus Training Set

A training dataset was formulated from a wide variety of coronavirus spike protein sequences from alpha, beta, gamma and delta coronaviruses and constituted isolates from many different animals. The total number of spike protein sequences in the training dataset was 2406, encompassing 511 sequences from Human CoV including examples of SARS-CoV-1 and MERS as well as SARS-CoV-2 (hCoV-19), 232 Bovine, 194 Noctilionine (Bat), 106 Porcine and several samples from other animals including camel, Chinese ferret-badger, hedgehog, dog, deer, avian and whale. Downloaded sequences were searched and cleaned to remove poorer quality and partial sequences and subunits resulting in a total of 2295 sequences. All the cleaned data was used in training and the model was evaluated by bioinformatic methods. BLASTP percentage match identities within the dataset had a mean of 79.7, median of 92.5 and an interquartile range of 37.2.

3 Results

3.1 Characteristics of DL Simulated Spike Proteins

To create predictions, the RNN is initially given a short seed protein sequence. The seed sequence can be passed as a random choice from previously sequenced spike proteins or formulated of random choices of amino acids starting with Methionine chosen by the python random library. In this study, given a seed text, the RNN was then able to provide sequences up to the full length of spike protein, a maximum length in the input dataset of 1582 amino acids with a mean length of 1324.4 amino acids. The maximum sequence identity that a simulated sequence achieved in BLAST matches against the training set was 100% sequence identity over 875 amino acids with a temperature scaling value of 1.0 (see below for details) or 100% identity over the full length of the protein with a temperature scaling value of 0.5. The lengths of all the synthesized proteins were fixed at 1588 amino acids.

For preliminary investigations, 100 DL synthesised spike protein sequences were collected. The RNN was initially provided with seed sequences of 16 amino acids chosen at random from the starts of the full dataset of spike proteins. The amino acid complement of the real and synthesized spike proteins in the datasets is as compared below in Fig. 3. Although the amino acid complements show some differences, there are significant similarities across the two datasets.

Sequence Matching
All 100 of the simulated sequences had a significant BLASTP match to Spike protein from one or more coronavirus sequences with BLAST searches of the query sequences against the entire non-redundant database (Fig. 4).

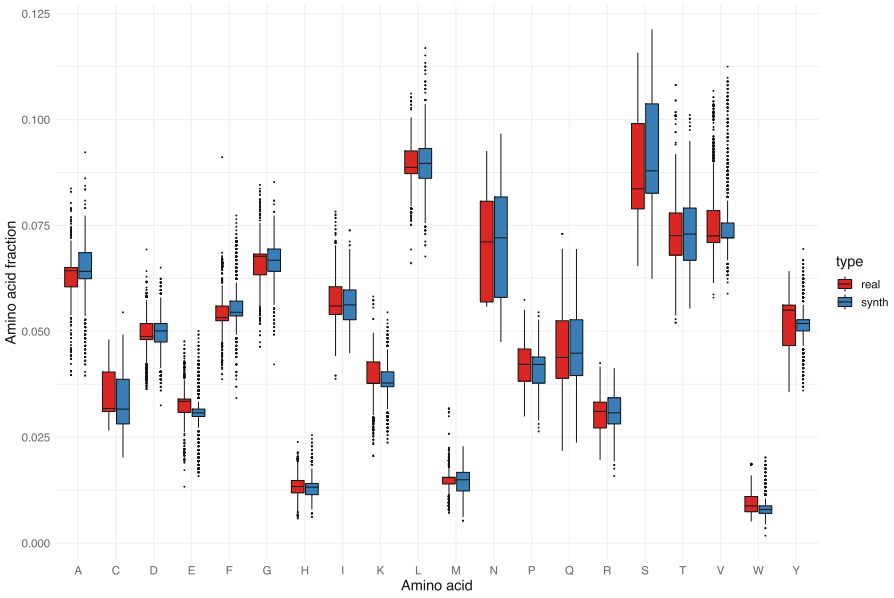

Fig. 3. Comparison of the amino acid composition of the real and simulated proteins. Boxplot graph showing the amino acid composition of each amino acid as a fraction of the protein sequence in both the real dataset (red) and the synthesized dataset (blue). The amino acid single letter code is shown on the X axis with the fraction of the amino acid in each sequence on the y axis as calculated by Biopython ProtParam module. The 'Real' training dataset comprised 2295 sequences in total with the 'Synth' simulated example dataset containing a matched number of samples. The difference between the Real and Synth amino acid composition datasets was not significant (Mann-Whitney test in R, p-value = 0.35).

```
Query 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS  60
        MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNS TRGVYYPDKVFRSSVLH TQDLFLPFFS
Sbjct 1 MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSSTRGVYYPDKVFRSSVLHLTQDLFLPFFS  60
```

Fig. 4. Partial BLAST alignment of Spike protein from a simulated query protein against Bat coronavirus RaTG13 [10]

The real spike protein training dataset was clustered and deduplicated resulting in 154 clusters of non-redundant sequences. The 100 simulated sequences were searched with BLASTP against the representative cluster sequences. Figure 5A shows that the best BLAST hits for the first set of simulated sequences covered several distinct clusters. There were several hits to MERS clusters, possibly due to a high representation of MERS sequences in the training set. A second set of 100 simulated sequences were generated that each had an identical seed text of 64 amino acids from the start of SARS-CoV-2 spike. Figure 5B shows that the equivalent BLAST hits on these sequences had a higher number of SARS-CoV matches, as well as Bat SARS-like sequences, although some samples still shared high identities with MERS sequences.

5A

5B

Fig. 5. BLAST matches of simulated sequences against known Spike proteins
Figure 5A shows BLAST searches of the original 100 DL synthesized sequences filtered for matches by length over 200bp and identity over 90%. 5B shows the second set of DL synthesized sequences which were all given identical seed text feeder sequence of 64 amino acids from the start of SARS-CoV-2. This graph is filtered for matches by length over 500 bp and identity over 90%. Highest length matches are represented by the largest diameter circle and darkest colour. However, large circles are generally of interest since the identity cut-off is high. Unfiltered data can be found at the GitHub site as described in the Data Availability section.

Pfam Domain Complements of Simulated Protein Sequences

Significant Pfam domain hits were uncovered on searching the query sequences with HMMER3 against the Pfam-A database. Searches of the synthesized proteins against Pfam_A.hmm database revealed Pfam domains that were expected within a coronavirus spike protein (below).

Table 1 Common Pfam domains and their counts identified within the original 100 simulated sequences which are also found in real Spike proteins, showing that all 100 sequences had C-terminal Spike domains. Other domains were identified in full Pfam-A however, the most common were Corona_S2, Spike_rec_bind and Spike_NTD. Database Pfam-A.SARS-CoV-2 refers to the April 2, 2020 update for SARS-CoV Pfam domains (Xfam Blog https://xfam.wordpress.com/2020/04/02/pfam-sars-cov-2-special-update/).

The resulting Pfam domains compared favourably with the most commonly found domains within the real training dataset of 2504 proteins which had 1781 domain counts of Spike_rec_bind, 2413 Corona_S2, 1052 Spike_NTD and 502 Corona_S1 (Coronavirus S1 glycoprotein domain) among others. According to Pfam Architectures, domain Corona_S2 is found in real spike proteins in the databases together with

Table 1. Pfam Domain Complements in the 100 simulated sequences.

Pfam domain	Pfam database	Full name	Count
Corona_S2	Pfam-A	Coronavirus S2 glycoprotein	100
Corona_S1	Pfam-A	Coronavirus S1 glycoprotein	13
Spike_rec_bind	Pfam-A	Spike receptor binding domain	81
Spike_NTD	Pfam-A	Spike glycoprotein N-terminal	53
CoV_NSP2_C	Pfam-A.SARS-CoV-2	Coronavirus replicase NSP2, C-terminus	6
CoV_S1_C	Pfam-A.SARS-CoV-2	Coronavirus Spike S1, C-terminus	75
bCoV_S1_RBD	Pfam-A.SARS-CoV-2	Betacoronavirus Spike S1, receptor-binding	82
bCoV_S1_N	Pfam-A.SARS-CoV-2	Betacoronavirus-like spike S1, N-terminus	88
CoV_S2	Pfam-A.SARS-CoV-2	Coronavirus Spike glycoprotein S2	100

either Corona_S1, Spike_NTD and Spike_rec_bind, or with just Spike_rec_bind, or with Spike_NTD and 2 x Spike_rec_bind or in some sequences as a standalone domain.

Prediction

During prediction, probabilities are generated for the next character in the sequence of the amino acid single letter alphabet. A parameter, known as temperature, can be used to scale the probabilities of the output distribution. If the temperature value is low, the model will be more confident on predictions which may produce more repetitive text. At a temperature of 0.5, the model was able to reach 100% identity over the full length of Spike SARS glycoprotein with the only differences being in the seed text. The purpose of this study is to provide sequences that are not identical to known sequences so we may find better use of a higher temperature value to provide more diverse text.

A second dataset sample of 100 synthesized sequences was formed by specifically using a seed text of 64 amino acids from the SARS-CoV-2 spike protein for each simulated sequence. When simulated sequences were clustered at the default 90% level of sequence identity, the result was 51 separate clusters in which Cluster 0 had 27 members ranging from 92%–100% identity which corresponded to SARS-CoV-1 type, Cluster 1 had 13 members of 97–100% identity which corresponded to Bat RaTG13/SARS-CoV-2 type, Cluster 38 had 3 members corresponding to MERS type, Cluster 2 had 3 members, Clusters 4, 8 and 19 each had two members, whilst each example of the rest of the dataset clustered separately. Hence, the seed text provided SARS-like hits in several but not all cases. Some sequences were definitively of interest to this study, such as a synthesized protein with 97% full length identity to a Bat beta-coronavirus sequence isolated from *Chaerephon plicata* in Yunnan in 2011 [11]. Further sequences of interest included those with high identity over stretches of the protein sequence to SARS-CoV-1 or SARS-CoV-2, particularly those including hybrid regions. Once the initial predicted protein is finished, the prediction commences a new protein again immediately if the maximum number of characters has not been reached. In some cases there were hybrid matches to parts of sequence from spike proteins in the dataset. A larger dataset of 1000 simulated sequences was generated with the same SARS-CoV seed text as previously.

These simulated sequences were clustered and clusters corresponding to SARS-CoV-1 and SARS-CoV-2 were aligned together with examples of the real spike proteins from SARS-CoV-1 and SARS-CoV-2. Multiple sequence alignments of the binding region indicate that residues important in human ACE2 recognition [12, 13] are broadly conserved across the simulated sequences.

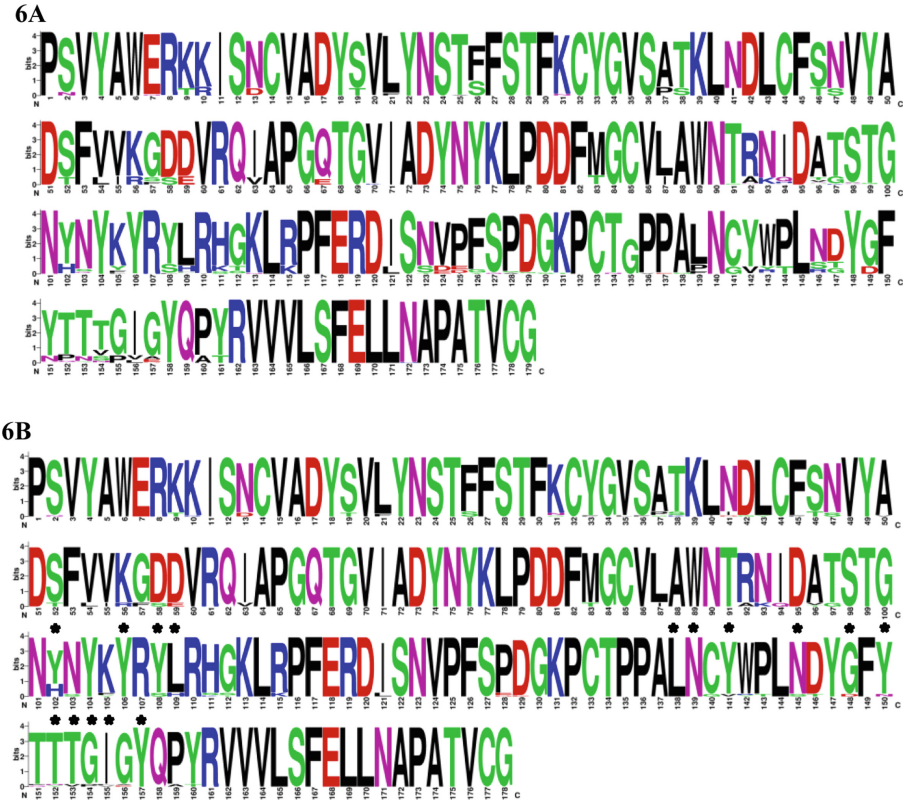

Fig. 6. Sequence Logos from multiple sequence alignments of the spike receptor binding domains (RBD) generated on Weblogo3 [14]. Each logo consists of stacks of symbols, one for each position in the sequence. Colours denote amino acid chemical properties. The height of the stack indicates the relative frequency of each amino acid whilst the width of the stack shows the fraction of symbols in the column (narrow = many gaps). Figure 6A shows the receptor binding domains of all the proteins in the real dataset that possessed a RBD matching that of the SARS-CoV and SARS-CoV-2 clusters in the dataset. Figure 6B shows the RBD of the synthetic dataset alignments that matched those clusters. The starred amino acids are contacting residues with human ACE2 receptor in the RBD of both SARS-CoV and SARS-CoV-2 [13].

4 Conclusions

This study used a comprehensive training set formulated from Coronavirus Spike protein sequences in DL neural networks to produce novel sequence from a short feeder seed text. Novel sequences shared features that can be searched with bioinformatics tools to provide highly significant BLAST and Pfam domain matches. That each of the sequences examined shared matches to Spike protein is exciting and warrants further consideration. Interestingly, in one example, the prediction query was able to correct an unknown amino acid (X) to a G that exactly matched other sequences of that type. It is trivial to generate high numbers of these synthesized sequences from the model, although in some cases the resultant sequence may not represent viable protein. In addition, the training set is only as comprehensive as the initial database, animal CoV sequences may exist elsewhere that are not represented here, and further unknown biases may exist.

Synthesised sequences may find a use in cases of data privacy such as generating synthetic patient data for studies as provided by Synthea (Standard Health Record Collaborative), or to generate further examples of poorly represented data for better statistical analysis. Further applications may include the generation of gene clusters of a particular type and evolution studies. Whilst the CoV model occasionally produced data with large-scale rearrangements, the results indicated that relatively simple neural networks can provide useful synthetic sequences with low compute requirements when trained on a curated database. The potential production of novel sequence by DL is exciting as a future strategy and warrants further consideration.

Data Availability. The model and source code are available at: https://github.com/LCrossman.

References

1. Organization WH: Consensus document on the epidemiology of severe acute respiratory syndrome (SARS). WHO/CDS/CSR/GAR/2003.11 (2003)
2. Zaki, A.M., Van Boheemen, S., Bestebroer, T.M., et al.: Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. (2012). https://doi.org/10.1056/NEJMoa1211721
3. Zhou, P., Fan, H., Lan, T., et al.: Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. Nature (2018). https://doi.org/10.1038/s41586-018-0010-9
4. Zhu, N., Zhang, D., Wang, W., et al.: A novel coronavirus from patients with pneumonia in China, 2019. N. Engl. J. Med. (2020). https://doi.org/10.1056/NEJMoa2001017
5. Goodsell, D.: Molecule of the Month SARS-CoV-2 Spike (2020). https://doi.org/10.2210/rcsb_pdb/mom_2020_6. http://pdb101.rcsb.org/motm/246. Accessed 14 June 2022
6. Li, F.: Structure, function, and evolution of coronavirus spike proteins. Annu. Rev. Virol. (2016). https://doi.org/10.1146/annurev-virology-110615-042301
7. Zhou, G., Zhao, Q.: Perspectives on therapeutic neutralizing antibodies against the Novel Coronavirus SARS-CoV-2. Int. J. Biol. Sci. (2020). https://doi.org/10.7150/ijbs.45123
8. Cho, K., Van Merriënboer, B., Gulcehre, C., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference (2014)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. (1997). https://doi.org/10.1162/neco.1997.9.8.1735

10. Zhou, P., Lou, Y.X., Wang, X.G., et al.: A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature (2020). https://doi.org/10.1038/s41586-020-2012-7
11. Wu, Z., Yang, L., Ren, X., et al.: ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. J. Infect. Dis. (2016). https://doi.org/10.1093/infdis/jiv476
12. Luan, J., Lu, Y., Jin, X., Zhang, L.: Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection. Biochem. Biophys. Res. Commun. (2020). https://doi.org/10.1016/j.bbrc.2020.03.047
13. Lan, J., Ge, J., Yu, J., et al.: Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature (2020). https://doi.org/10.1038/s41586-020-2180-5
14. Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E.: WebLogo: a sequence logo generator. Genome Res. (2004). https://doi.org/10.1101/gr.849004