Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 152 (2023) 106264

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution

Binbin Zhou [a,d,1], Hang Zhou [a,b,1], Xue Zhang [c], Xiaobin Xu [c], Yi Chai [g], Zengwei Zheng [a,d], Alex Chichung Kot [h], Zhan Zhou [c,e,f,*]

[a] Department of Computer Science and Computing, Zhejiang University City College, No. 48 Huzhou Street, Hangzhou, 310015, China
[b] College of Computer Science and Technology, Zhejiang University, Hangzhou, 310027, China
[c] Innovation Institute for Artificial Intelligence in Medicine and Zhejiang Provincial Key Laboratory of Anti-Cancer Drug Research, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, 310058, China
[d] Industry Brain Institute, Zhejiang University City College, Hangzhou, 310015, China
[e] The Fourth Affiliated Hospital, Zhejiang University School of Medicine, Yiwu, 322000, China
[f] Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou, 310058, China
[g] ZJU-UoE Institute, Zhejiang University, Haining, 314400, China
[h] School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore

ARTICLE INFO

Keywords:
SARS-CoV-2
Viral evolution
Natural language processing
Transformer-based method
Phylogenetic tree
Mutation prediction

ABSTRACT

The widespread of SARS-CoV-2 presents a significant threat to human society, as well as public health and economic development. Extensive efforts have been undertaken to battle against the pandemic, whereas effective approaches such as vaccination would be weakened by the continuous mutations, leading to considerable attention being attracted to the mutation prediction. However, most previous studies lack attention to phylogenetics. In this paper, we propose a novel and effective model TEMPO for predicting the mutation of SARS-CoV-2 evolution. Specifically, we design a phylogenetic tree-based sampling method to generate sequence evolution data. Then, a transformer-based model is presented for the site mutation prediction after learning the high-level representation of these sequence data. We conduct experiments to verify the effectiveness of TEMPO, leveraging a large-scale SARS-CoV- 2 dataset. Experimental results show that TEMPO is effective for mutation prediction of SARS- CoV-2 evolution and outperforms several state-of-the-art baseline methods. We further perform mutation prediction experiments of other infectious viruses, to explore the feasibility and robustness of TEMPO, and experimental results verify its superiority. The codes and datasets are freely available at https://github.com/ZJUDataIntelligence/TEMPO.

1. Introduction

Since the first report of coronavirus disease (COVID-19) in late December 2019, which was caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), more than 600 million infection cases have been reported from more than 200 countries or regions around the world, according to the World Health Organization (WHO) [1]. The widespread of SARS-CoV-2 presents a significant threat to human life, as well as public health and economic development, over 6.48 million lives have been lost in the COVID-19 outbreak and unquantifiable economic losses up to July 2022 [2]. Extensive efforts have been put out and implemented to battle against the pandemic, including the development of pharmaceutical interventions, such as drugs, antibodies, and vaccines, as well as non-pharmaceutical approaches, such as quarantine and keeping a physical distance [3–8]. Vaccination has been widely used as a public anti-epidemic strategy since it has been proven to be a promising approach. However, the efficacy of current vaccines would be diminished by the emerging SARS-CoV-2 variants. With the spread of the pandemic, SARS-CoV-2 has continued to evolve and mutate, with an increasing number of mutations

* Corresponding author. Innovation Institute for Artificial Intelligence in Medicine and Zhejiang Provincial Key Laboratory of Anti-Cancer Drug Research, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, 310058, China.
E-mail addresses: bbzhou@zucc.edu.cn (B. Zhou), hangz@zju.edu.cn (H. Zhou), 22119130@zju.edu.cn (X. Zhang), 3190104201@zju.edu.cn (X. Xu), ychai@u.nus.edu (Y. Chai), zhengzw@zucc.edu.cn (Z. Zheng), eackot@ntu.edu.sg (A.C. Kot), zhanzhou@zju.edu.cn (Z. Zhou).
[1] These authors have contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2022.106264
Received 18 September 2022; Received in revised form 16 October 2022; Accepted 30 October 2022
Available online 14 December 2022

Table 1
The number of the available SARS-CoV-2 spike protein sequences from 2020.01 to 2022.02.

Month	2020.01	2020.02	2020.03	2020.04	2020.05	2020.06	2020.07	2020.08	2020.09
	756	1800	56,173	57,978	29,898	33,283	42,785	43,198	46,589
Month	2020.10	2020.11	2020.12	2021.01	2021.02	2021.03	2021.04	2021.05	2021.06
	79,322	108,871	149,285	253,874	276,977	431,163	443,089	343,410	284,341
Month	2021.07	2021.08	2021.09	2021.10	2021.11	2021.12	2022.01	2022.02	
	547,335	854,129	706,695	653,010	798,758	625,360	288,256	232,881	

arising. According to the RCoV19 dataset [9], the variation frequency of the medium value of SARS-CoV-2 reached 0.264% and more than 1851 lineages [10] have emerged up to July 2022. The mutations will make it more difficult for the virus to be identified by human antibody-mediated vaccinations, which will either render the vaccines ineffective or inhibit diagnostic detection [11–15].

Tracking the evolution of SARS-CoV-2 could provide a thorough insight into viral evolution dynamics and early detection of variants of concern (VOCs). Virus evolution is primarily driven by mutations, both genetically and antigenically. Bai et al. successfully forecasted that some certain mutations of the spike protein at N501 may result in a stronger binding of angiotensin converting enzyme 2 (ACE2), prior to the emergence of the UK mutant (SARS-CoV-2 VOC 202012/01) on December 1, 2020 [16]. Starr et al. investigated the impact of amino acid mutations to the receptor binding domain (RBD) on the expression of folded protein as well as its affinity for ACE2, and showed that the N501Y mutation in the Alpha variant strengthens binding to ACE2 [17]. Tegally et al. presented a new lineage of the Beta variant of SARS-CoV-2 with three residues in the binding site (K417 N, E484K and N501Y), which has spread incredibly quickly and become one dominant lineage in a few weeks [18]. Similar to the B.1.351 variant, the Gamma variant, which possesses independent K417T, E484K and N501Y mutations, is quickly spreading over regions [12]. The fast spread of the Delta variant (B.1.617.2) that exploded in early 2021 was investigated, and the results revealed that it was caused by the evasion of antibodies, growing activity, and higher transmissibility [19,20].

Accurate and efficient prediction of genetic mutations of the SARS-CoV-2 has attracted significant attention, which could contribute to better identifying vulnerabilities for antibody-based treatments, vaccines, and diagnostics, as well as increasing the window of opportunity for developing proactive responses [21–24]. With the advent of high-throughput technologies and the availability of massive sequence data, numerous computational models for viral evolution have been developed in recent years. Xia et al. designed a statistical mutual information-based approach to calculate the variance correlations of sites, and used the most recently mutated sites to infer the future most probable mutated sites [25]. Bai et al. evaluated the free energy change of different kinds of single site or combinedsite mutations and predicted possible mutation sites [26]. Yin et al. modeled the temporal sequence of sites and employed a long short-term memory (LSTM) model with an attention mechanism to predict mutations at each single target site [27]. Maher et al. predicted the possible mutations in SARS- CoV-2 contributing to future variants by using a variety of methods, such as a bidirectional LSTM model, to evaluate mutational effects, and identified the key biological drivers of intrapandemic evolution [28]. In addition, using natural language models as a powerful computational model to handle biological sequence data has become a popular idea due to its rapid development and effectiveness. As an example, Hie et al. proposed a self-supervised learning approach to learn the sequence representation of proteins to maximize the immune escape capability while ensuring the fitness of protein sequences [29]. However, despite a large number of studies using computational models to predict viral evolution, most of the previous work does not take into account the phylogeny of viral evolution, which significantly reflects the development process of variants. In addition, due to the complexity of the evolutionary process and the diversity of evolutionary lengths, it is necessary to design models with more expressive power and computational efficiency. Therefore, we choose to design methods for this mutation prediction problem based on the transformer model, within which the attention architecture is naturally conducive to better capturing long-range dependencies and facilitating large-scale parallel computation.

In this paper, we propose a transformer-based mutation prediction framework (TEMPO) for SARS-CoV-2 evolution. TEMPO can learn a high-level representation of historical prior sequence data that are constructed from the phylogenetic tree (PT) structure data of SARS-CoV-2, and afterwards predict the mutation probability of sites. More specifically, we first design a systematic PT-based sampling method to generate the sequence of viral amino acid sequences combined with temporal information, in which the temporal nature of viral evolution can also be portrayed. Second, we employ an embedding method ProtVec [30] for the residue representation. After that, a transformer structure is introduced to encode the embedded sequence to extract and learn complex correlations, and a fully connected layer is performed for the final site variants prediction. In addition, we devise an architecture to predict the mutations at each single site. We conduct experiments to verify the effectiveness of our proposed model TEMPO on the mutation prediction of SARS-CoV-2 evolution leveraging the SARS-CoV-2 dataset. Experimental results demonstrate that TEMPO is effective and outperforms several state-of-the-art baseline methods. We further explore the feasibility and robustness of TEMPO by performing mutation prediction experiments of other infectious viruses leveraging three influenza datasets, including the H1N1, H3N2 and H5N1 subtypes. The results indicate that our model is not only effective for the mutation prediction of SARS-CoV-2, but also informative for the evolutionary analysis of a broader range of viruses. We anticipate that the proposed TEMPO can provide significant insight into the evolution of SARS-CoV-2 and contribute to the early detection of VOCs in the next stage.

2. Materials and methods

2.1. Datasets

2.1.1. Data collection

For SARS-CoV-2, we collected the amino acid sequence data of its spike protein from the GISAID database [31] and its phylogenetic tree data from the RCoV19 database [9]. The phylogenetic tree was generated by Pangolin software (Pangolin:4.1.2, PangoLearn: 2022-04-09). Table 1 presents the statistical information of SARS-CoV-2 spike protein sequence data and a total of 7389216 sequences were used after data cleaning. Mutation prediction in Spike proteins is our target, since Spike is the focus of antibody-mediated immunity and is the principal antigen in existing vaccines [32]. In addition, HA protein sequence data of influenza subtypes H1N1, H3N2 and H5N1 were also collected for additional experimental validation, following Yin's work [27]. Since these datasets contain series with different time spans, the SARS-CoV-2 dataset is split by month, while the three influenza virus datasets are by year.

2.1.2. Data preprocessing

Because our method performs single site mutation prediction, sequence alignment is required to ensure the consistency of site numbering in different samples. Multiple sequence alignment (MSA) was performed on the protein sequences using MAFFT [33]. The SARS-CoV-2 dataset contains more than 8 million sequences in total, so

it is impractical to perform multiple sequences directly on the entire dataset. Due to the high similarity between the SARS-CoV-2 sequences, we compared each sequence in the SARS-CoV-2 dataset with the standard reference sequence pairwise rather than a direct multiple sequence comparison. To facilitate the subsequent phylogenetic-tree-based sampling method, some additional information, such as lineage and sequence quality, is added to each sequence from the metadata based on the unique sequence ID. Sequences without these additional information will be excluded.

Algorithm 1. PT-based Sampling Method.

Algorithm 1 PT-based Sampling Method

Input: Phylogenetic tree PT, original dataset (consisting of separate sequences) S, sample numbers of each path N;
Output: Training dataset (consisting of historical sequences of separate sequences) D;

```
D = { }
Paths = DFS (PT)
for ∀path ∈ Paths do
    for i = 1 to N do
        Initialize d as an empty list
        for ∀node ∈ path(PT) do
            X = {x|x.lineage = node.lineage}
            Randomly select x from X
            Insert x at the end of d
        end for
        D.add (d)
    end for
end for
```

2.2. PT-based sampling method

To predict virus mutation based on evolutionary information, the first step is to generate evolutionary historical sequences from discrete protein sequences in the dataset. More explicitly, each historical sequence is composed of multiple protein sequences. The historical sequences can be used to reflect the evolution of the virus to better predict future mutation trends. A simple approach to obtain historical sequences is sampling from the dataset sequentially in chronological order based on the submission time of the sequences. However, the chronological order of sequence submission does not always coincide with the evolutionary order. For example, some of the most recently submitted sequence data may belong to a lineage that is early in the evolution of the virus, while some relatively earlier submitted sequences may also belong to a lineage that has just emerged at that time and they are at the further back in the evolutionary order, which we should pay more attention to.

Therefore, we propose a phylogenetic tree (PT) based method sequentially sampling viral sequences to construct historical sequences that better reflect the virus evolutionary order, as shown in Algorithm 1. A phylogenetic tree is a branching diagram that depicts the evolution of various species or lineages from a common ancestor. This can help understand what occurred throughout evolution. Specifically, the phylogenetic tree data were obtained from the RCoV19 database [9] which is based on Phylogenetic Assignment of Named Global Outbreak Lineages (PANGO). First, a depth-first search (DFS) method is performed on the evolutionary tree to retrieve the set of evolutionary paths, where each evolutionary path corresponds to a path from the root node to a leaf node of the phylogenetic tree. Second, we generate multiple samples in each evolutionary path. These samples can be treated as historical sequences since they reflect the historical evolution process. When generating a single sample, each node on the path is selected in turn according to the order from the root node to the leaf node, and the amino acid sequence with the same lineage is discovered and added to the sample. The generation process of one sample would be completed when it reaches the last leaf node on the evolutionary path.

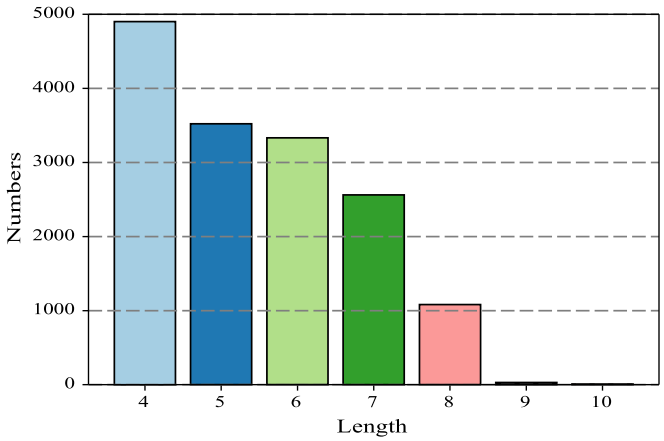

Fig. 1. Length distribution of the sampled historical sequences (i.e., the number of nodes in the corresponding evolution path).

The length distribution of sampled historical sequences is shown in Fig. 1. The results show that the path lengths of the generated sequences are all less than 10, and most are concentrated between 4 and 8, which indicates that the phylogenetic tree is wide and shallow. To avoid evolutionary paths that are too short to reflect virus evolutionary characteristics, we selected evolutionary paths of length 5 to 8 for sampling. We also discard sequences whose evolutionary paths are longer than 8 because they are too few to construct a valid dataset which may also cause overfitting.

2.3. The proposed transformer-based model TEMPO

With these sampled historical sequence data, we develop a transformer-based model TEMPO for mutation prediction of SARS-CoV-2 evolution. The framework of this model has been presented in Fig. 2. We would elaborate on it as below.

2.3.1. Sequence encoding

Sequence sampling is simply the process of obtaining historical series data of amino acid sequences from a dataset. To perform mutation prediction using machine learning models, amino acid sequences (encoded with character 'ACDE … ') need to be encoded as vectors of real numbers, and the protein language pre-trained model ProtVec [30] is used to complete this process.

The specific steps are shown in Fig. 3. First, we split the whole sequence into subsequences and generate the embedding matrix of 3-g based on ProtVec [30]. Inspired by Tempel [27], we break these sequences into shifted overlapping residues in the window of 3 shown in Fig. 3. For example, in SARS-CoV-2, each spike protein sequence is depicted as 1273 lists of 3-g that are embedded in a 1273*100 dimensional vector space based on ProtVec [30], where a 3-g is represented by a 100-dimension vector space. The 'unknown' vector from ProtVec will be assigned to denote the subsequence if it contains '-' at any positions. To predict the mutation for each site, we utilize three overlapping 3-g, shown in Fig. 3, and focus on the center position as the target site. The three overlapping 3-g would be represented as the summation of the individual 3-g embedding vector. Therefore, each training case incorporates n sequential 3-3-g, embedded in 100*n dimensions, where n is the length of the evolutionary path. An example of generating the representation of historical sequences can be found in Supplementary

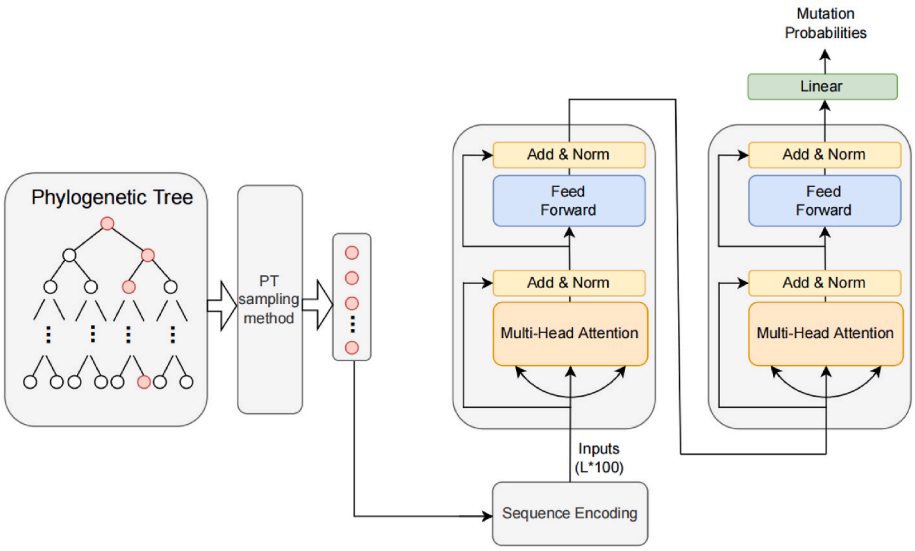

Fig. 2. The framework of the proposed model TEMPO for mutation prediction of SARS-CoV-2. L represents the evolutionary path length.

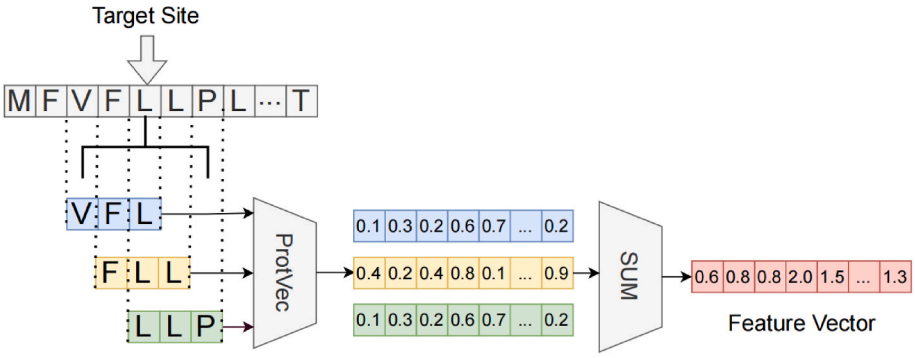

Fig. 3. Sequence encoding process. Each 3-g is represented by a 100 dimension vector. Representations of three 3-g around the target site are summed to generate the final feature vector.

Material.

2.3.2. *Transformer encoder model*

After sequence encoding, we use a transformer-based architecture to learn the evolutionary features of each target site in the entire input historical sequences, as presented in Fig. 2. Due to the complexity of the evolutionary process and the diversity of evolutionary lengths, models with more expressive power and computational efficiency are needed. Transformer is one popular sequence model architecture in the field of natural language processing recently. Compared to RNN, its internal attention mechanism is inherently computationally parallelizable and has better representation capabilities for especially long sequences, which is very suitable for virus evolution prediction tasks on large-scale datasets with evolutionary paths of various lengths. Correspondingly, in the task of virus mutation prediction, the evolutionary path of the target site in historical sequences can be viewed as a sentence, where each amino acid residue at the target site is considered as a word or a token. So predicting whether mutation will occur along this evolutionary path at the target site is transformed into a binary classification problem for a sentence.

Unlike classical time-series models such as RNN and LSTM, it is generally believed that transformer outperforms those methods due to two properties itself. First, it uses attention mechanism to adaptively learn the weights (i.e., the degree of influence on the features of next layer) of each element of the sequence for the feature embedding. Second, the pure attention architecture is designed to better capture long-range dependencies and facilitate large-scale parallel computation, which makes it easier to train larger parametric models in less time and improve the expressiveness of the model.

The original transformer is designed in the encoder-decoder architecture for machine translation tasks. In our task, since we only need to learn the features of the input sequence and make prediction based on the sequence feature, there is no need to generate a sequence in a transduction manner as translation tasks. Therefore, only the encoder part of the transformer architecture is reserved in our model. Specifically, the model input is a vector of L*100 dimensions, where L is the length of evolutionary path (referred as sentence later) in a historical sequences sample. The output obtained after the last layer of the transformer encoder is a vector of L*d dimensions, where d is the hidden layer dimension, a hyperparameter. These L*d-dimensional vectors can be viewed as the corresponding feature representations of each of the L target sites (referred to as tokens later) in the input historical sequence learned by the model.

Since we view the mutation prediction task as a sentence classification problem, we need to obtain the representation of the whole sentence from the representation of these L "words". Two common solutions are tried here: one is to take the average of representations of these L words as the representation of the whole sentence, and the other is to select only representation of the last word. It is worth mentioning that although the second approach only employs the representation corresponding to the last word, this representation also contains information about the whole sentence because the attention mechanism takes into

account all the words in the sentence during the computation.

The experimental results show that the latter method is more effective. We believe that the reason for this phenomenon may be that the attention mechanism itself assigns a weight to each word in the sentence. When using the mean value of the representation of each word as the sentence representation, the operation of averaging suppresses the variability in the degree of contribution of each word to the sentence representation, thus making it more difficult to train the model to obtain a better sentence representation, whereas when only the representation of the last word is taken as the sentence representation, the sentence representation is computed by the pure attention mechanism, thus avoiding the problem of averaging suppression.

Finally, the representation of the sentence is fed into a fully connected neural network, and the final prediction is obtained through a Softmax layer. The cross-entropy loss function is used to train the parameters in the network because cross-entropy is commonly used to measure the difference between two probability distributions and is therefore well suited as an objective function for the binary classification task. In addition, there are some details worth mentioning in training processes. In order to train the network in a supervised manner, it is necessary to give a label to each input sequence. We generate labels based on whether the last two protein sequences in the input historical sequence (i.e., the last two nodes on the evolutionary path) have the same amino acids at the target site. When the amino acids of the last two protein sequences at the target site cannot be determined (taking the value X or missing), this sample will be ignored and excluded from the training process.

2.4. Experiments

2.4.1. Dataset settings

We collected in excess of 8 million SARS-CoV-2 sequences from January 2020 to February 2022 from the GISAID database, and generated a total of 15,200 samples with 5 samples for each evolutionary path to evaluate the effectiveness of our method. Specifically, we first preprocess the dataset to filter abnormal sequences and add additional information needed, and then use the proposed PT-based sampling method to generate training samples.

2.4.2. Baselines

Two types of baselines are set up for prediction, as follows:

2.4.2.1. Traditional machine learning methods.

- SVM [34]: Support vector machine (SVM) is a generalized linear classifier for binary classification. The embedding of the target site at the penultimate node on the evolution path is fed to predict mutation at the leaf node.

- LR [35]: Logistic regression (LR) is a traditional machine learning method which maps the input embedding features at the penultimate node to a scalar in (0,1) with the logistic function which corresponding to the classification probability.

- RF [36]: Random forest (RF) is an ensemble learning method which uses multiple decision trees to do the classification task. The input and output settings are the same as SVM and LR.

2.4.2.2. RNN-based methods.

- RNN [37]: Recurrent neural network (RNN) is a specialized neural network for sequences, which allows variable length sequences as input and output. Specifically, the sequence of target node embedding with the sequence length N is used as the model input, where N is the length of the corresponding evolution path. And the output is a 2-dimensional vector which indicates the probability of mutation or not.

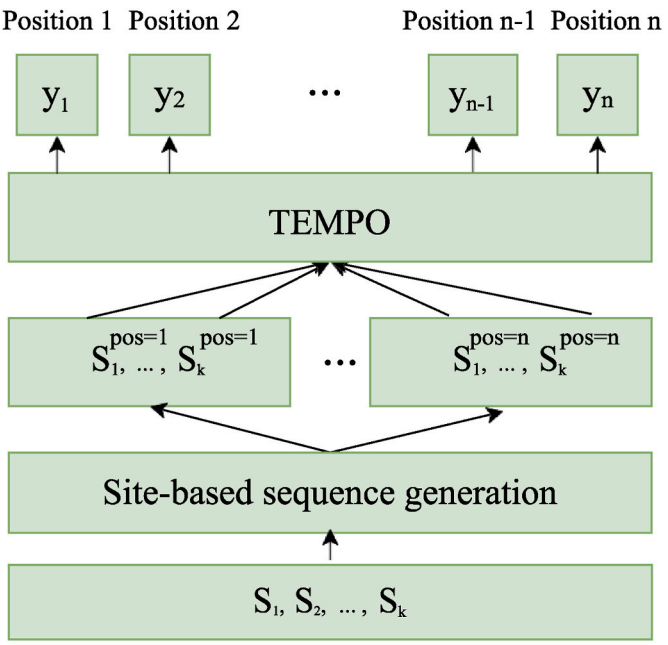

Fig. 4. The workflow of mutation prediction of SARS-CoV-2 at specific sites using TEMPO.

- LSTM [38]: LSTM is an RNN-type network which can capture features of longer sequences with several well designed gating mechanism.

- Tempel [27]: Tempel is a LSTM-based model with an attention mechanism to capture more complex correlations in the input sequence.

2.4.3. Implementation details

All the approaches are performed with Scikit-learn [39] and Pytorch [40]. Note that the path from the root node to the leaf node in the phylogenetic tree is denoted as the *evolutionary path*. In the implementation, we use all the nodes without leaf nodes on the *evolutionary path* to predict the mutation of the leaf node. Specifically, we divide the dataset into a training data and a testing data according to a ratio of 4:1. In the training dataset, we generate the evolutionary historical sequence data based on the phylogenetic tree. In order to forecast the ultimate mutation, we first remove the leaf nodes from the phylogenetic tree and use the data from all other nodes, aside from the leaf nodes, as input for the proposed transformer-based model. The training objective is to minimize cross-entropy loss. We train the model for 100 epochs to achieve the convergence. With the trained model, we input the testing data without the leaf nodes, to predict whether the leaf nodes would witness a mutation.

For all RNN-based models, we apply stochastic gradient descent with a minimum batch size of 256 for optimization. The learning rate is 0.001 with 128 hidden units in the encoder. We use cross-entropy as the objective function. The strategy with a drop-out of 0.5 is carried out and all the models are fit for 200 training epochs. We implement all baselines and our model on the environment with one AMD EPYC 7502P CPU @ 3.35GHZ and NVIDIA RTX3090 24 GB card.

2.4.4. Evaluation metrics

We use five commonly used metrics, i.e., Accuracy, Precision, Recall, F-score and Matthews correlation coefficient (MCC), for classification model performance evaluation, among which the most important

Table 2
Experiment results on the SARS-CoV-2 dataset.

Method	Accuracy	Precision	Recall	F-score	MCC
SVM	0.530	0.519	0.588	0.551	0.063
LR	0.542	0.530	0.575	0.552	0.085
RF	0.544	0.534	0.561	0.547	0.089
RNN	0.609	0.581	0.720	0.643	0.226
LSTM	0.648	0.619	0.731	0.670	0.302
Tempel	0.648	0.618	**0.743**	**0.675**	0.305
TEMPO	**0.655**	**0.658**	0.614	0.636	**0.309**

metrics are Accuracy, F- score and MCC. The detailed calculation formula for these metrics are shown in the Supplementary Material. Overall, a higher accuracy indicates a better precision rate of the prediction while a higher F-score value indicates a better trade-off of accuracy and completeness for positive sample prediction. The MCC is generally considered a more balanced indicator and can be applied even when the sample of the two categories differs significantly (i.e., category imbalance), which matches our scenario well because mutation will not occur in most cases.

3. Results and discussion

3.1. The workflow of TEMPO

We present a workflow, as shown in Fig. 4, to demonstrate the implementation steps of our method for mutations prediction at specific individual sites. Our objective is to predict whether there will be mutations in each site at the next stage of the evolution process, which is represented by the bottom leaf node of the phylogenetic tree as shown Fig. 2. The sampled historical residue information in sequence following the phylogenetic tree structure is mapped as the path from the root node to the penultimate node. Given historical sequences S_1 to S_k generated by the PT-based sampling method, we produce historical embedding sequences $S_1^{pos=i}$ to $S_k^{pos=i}$ for each target site position i. Then we feed all the samples into the TEMPO model to obtain final mutation prediction results for each target site.

3.2. Performance evaluation of TEMPO for SARS-CoV-2 mutation prediction

To evaluate the effectiveness of the proposed model, we conduct comparison experiments on the mutation prediction between our model TEMPO and various baseline methods, using a large-scale SARS-CoV-2 dataset. The experimental results are shown in Table 2. All the results are averaged over 10 random trails with a fixed random seed. The experimental

results show that TEMPO achieves the best prediction performance in terms of multiple evaluation metrics, including accuracy, precision, and MCC, with an improvement of 1.1%, 6.5%, 1.3% compared with baseline methods. This can be attributed to the attention architecture of the transformer encoder layer in TEMPO, which is able to capture the complicated correlations between the lineages throughout the evolution path and learn the importance of the impact of each lineage on the final mutation probability. We can observe that TEMPO has a lower recall performance, which may be attributed to the overfitting issue with more parameters in the transformer layers. In addition, SVM has the worst prediction performance. The possible reason may be that this baseline method is not feasible for effectively learning time-series information. Moreover, we observe that TEMPO and other RNN-based methods outperform other baselines significantly. This may be because non-sequential models have difficulty utilizing the history of the sequence evolution, which in turn proves that the evolution path of the target lineage plays an important role in the mutation prediction. This is determined by the strong correlations between the ancestor node and the child nodes on the phylogenetic tree, which can be captured effectively by the proposed PT-based sampling method. Among all the sequential models, Tempel and LSTM obtain comparable prediction performance, which can be explained by the similarity of their whole architecture.

Table 3
Experiment results on influenza datasets.

Datasets	Method	Accuracy	Precision	Recall	F-score	MCC
H1N1	LR	0.823	0.685	0.418	0.519	0.438
	RF	0.904	**0.805**	0.764	0.784	0.723
	RNN	0.897	0.773	0.774	0.774	0.707
	LSTM	0.902	0.799	0.761	0.780	0.717
	Tempel	0.902	0.798	0.763	0.780	0.717
	TEMPO	**0.905**	0.803	**0.774**	**0.788**	**0.727**
H3N2	LR	0.938	0.47	0.043	0.079	0.128
	RF	0.962	0.812	0.504	0.622	0.623
	RNN	0.953	0.709	0.408	0.518	0.516
	LSTM	0.961	0.782	0.515	0.621	0.616
	Tempel	0.961	0.780	**0.521**	0.625	0.619
	TEMPO	**0.963**	**0.826**	0.510	**0.631**	**0.633**
H5N1	LR	0.987	0.872	0.210	0.338	0.424
	RF	0.989	0.826	0.426	0.562	0.589
	RNN	0.986	**0.971**	0.105	0.189	0.317
	LSTM	0.990	0.870	0.414	0.561	0.596
	Tempel	0.990	0.895	0.395	0.548	0.591
	TEMPO	**0.990**	0.863	**0.429**	**0.573**	**0.605**

3.3. Performance evaluation of TEMPO for influenza mutation prediction

We further explore the feasibility and robustness of our proposed TEMPO model by performing mutation prediction experiments with other infectious viruses. Three influenza datasets are used, including influenza subtypes H1N1, H3N2 and H5N1. These datasets are downloaded from Influenza Virus Resource [41], containing the full-length HA sequences between 1991 and 2016. Finally 161,000, 132,000 and 102,000 sequential samples of H1N1, H3N2 and H5N1 were used for the experiment, respectively.

Note that we do not take the SVM method into consideration, owing to its significant time complexity on the large datasets. It should also be noted that when using the influenza dataset for virus evolution prediction, we did not use a PT-based sampling method, but a time-sequence-based sampling method due to the lack of relevant influenza phylogenetic tree data. Experiments on these influenza datasets are designed to verify the generalization and robustness of TEMPO, so the sampling method used to generate the dataset is not the main concern here. Experimental results are demonstrated in Table 3. From the table, we find that TEMPO outperforms these baseline methods consistently, in terms of various evaluation metrics, including Accuracy, F-score, and MCC. We also observe that TEMPO obtains comparable mutation prediction performance in terms of Precision and Recall. This reflects the good generalization ability of our model. This TEMPO model can be used as a general framework for mutation prediction on various types of viruses, including SARS-CoV-2 and these influenza viruses.

3.4. Parameter experiments

To study the impact of different hyperparameter values on the mutation prediction performance, we analyze the performance of models on the SARS-CoV-2 dataset by varying three significant hyperparameters, including the number of encoder layers, number of attention heads and learning rate. The hyperparameter study results are depicted in Fig. 5.

From the figure, we observe that the best prediction performance is obtained when the number of encoder layers equals to 2. An appropriate value of the number of encoder layers improves the model's ability to effectively capture evolutionary historical information. We also choose

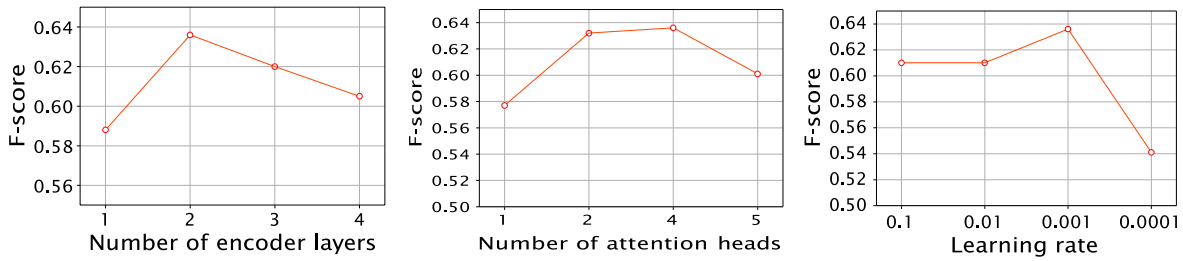

Fig. 5. Hyperparameter study of TEMPO.

Table 4
A list of high-probability ($p \geq 0.5$) predicted mutation sites in descending order.

Site	Prediction probability	Variant after February 2022
796	1.00	D796Y
452	0.94	L452R, L452Q, **L452M**
950	0.82	–
501	0.80	N501Y
484	0.78	E484A
859	0.78	–
681	0.75	P681H
215	0.75	**D215E**
144	0.69	–
936	0.67	–
156	0.67	–
142	0.62	G142D
222	0.60	–
19	0.59	T19I
158	0.57	–
157	0.56	F157L
478	0.54	T478K
145	0.50	**TH145–146-**
477	0.50	S477 N

different values of the number of attention heads n from 1 to 5 in the experiments, and find that the best F-score result is obtained when $n = 4$. When $n < 4$, the performance worsens due to the lack of representation ability. In addition, we search for the learning rate from 0.0001 to 0.1, and confirm the value 0.001 with the best performance in our model.

3.5. *Mutation prediction results*

We utilize all the data available up to February 2022 to predict the future mutation. Specifically, using all the data, we generate sampled sequence data and input it into the TEMPO model to forecast future mutations, following the same process as for the prediction framework described above. It should be noted that TEMPO only predicts the mutation probability of the target site and does not predict the specific type of mutation. We sorted the mutation prediction results at certain sites with more than 6 sampled sequence data, and presented a list of high-probability ($p \geq 0.5$) predicted mutation sites, as shown in Table 4. We compared the prediction results of TEMPO with the actual mutations according to the RCoV19 database v4.0 [9] and found that TEMPO can effectively predict mutations at certain sites where new variants arise. These new mutations have been illustrated in the bond font. For instance, according to our model, site 215 has a 0.75 chance to get a new mutation in the near future. Later real data documented in the RCoV19 database confirm the variants.

In addition, we can see that TEMPO can be used to predict mutations at sites that have not yet emerged, with 22 successfully predicted mutations among all the 39 newly emerged mutations, as shown in the Supplementary Material. Note that some mutations are predicted with few sampled sequence data. For example, site 259 has not undergone a mutation since February 2022, and according to our model, it has a probability of 0.75 to witness a mutation soon, which has been confirmed by the RCoV19 database. This is a challenging task, and we believe our model provides a promising step toward predicting new mutations.

4. Conclusion

Mutation prediction for SARS-CoV-2 evolution is a challenging and essential task. In this paper, we aim to predict future mutations in sites using historical spike protein data. We designed a phylogenetic tree-based sampling method to generate sequence data with full consideration of the tree-structure of SARS-CoV-2 evolution data. We propose a novel transformer-based model TEMPO that can fully utilize the prior sequence information and effectively learn high-level representations to enhance prediction performance. Experimental results on a large-scale SARS-CoV-2 dataset prove the effectiveness of our model for mutation prediction of SARS-CoV-2 evolution. To explore the feasibility and robustness of this model, we conducted further experiments on mutation prediction of other infectious viruses, leveraging three influenza datasets. The results verify the superiority of our proposed model.

Despite having obtained many research results, there still remain limitations. First, we did not take the phylogenetic tree construction method into consideration due to the construction computation difficulty in the large SARS-CoV-2 dataset. Thus, one open future work is to investigate the sensitivity of phylogenetic uncertainty when phylogenetic trees constructed by various methods are available. Second, we did not consider the global information of protein sequences. One of the future works will be to incorporate global information of proteins and investigate the effects on mutation prediction.

Data availability statement

The datasets and codes in this study can be found in the online repository https://github.com/ZJUDataIntelligence/TEMPO.

Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 62102349), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LDT23H19011H19), and the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China (Grant No. LHDMZ22H300002).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge the Supercomputing Center of Zhejiang University City College, the Information Technology Center and State Key Lab of CAD&CG of Zhejiang University, and Alibaba Cloud for the support of the advanced computing resources.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106264.

References

[1] WHO, WHO coronavirus (COVID-19) dashboard. https://covid19. who.int, 2022. (Accessed 26 July 2022).

[2] WHO, Coronavirus death toll. https://www.worldometers.info/coronavirus/coronavirus-death-toll, 2022. (Accessed 26 July 2022).

[3] Kizzmekia S. Corbett, Darin K. Edwards, Sarah R. Leist, Olubukola M. Abiona, Seyhan Boyoglu-Barnum, Rebecca A. Gillespie, Sunny Himansu, Alexandra Schäfer, Cynthia T. Ziwawo, Anthony T. DiPiazza, et al., Sars-cov-2 mrna vaccine design enabled by prototype pathogen preparedness, Nature (Lond.) 586 (7830) (2020) 567–571.

[4] Fatima Amanat, Florian Krammer, Sars-cov-2 vaccines: status report, Immunity 52 (4) (2020) 583–589.

[5] Wenhao Dai, Bing Zhang, Xia-Ming Jiang, Haixia Su, Jian Li, Yao Zhao, Xiong Xie, Zhenming Jin, Jingjing Peng, Fengjiang Liu, et al., Structure-based design of antiviral drug candidates targeting the sars- cov-2 main protease, Science 368 (6497) (2020) 1331–1335.

[6] Alicia T. Widge, Nadine G. Rouphael, Lisa A. Jackson, Evan J. Anderson, Paul C. Roberts, Mamodikoe Makhene, James D. Chappell, Mark R. Denison, Laura J. Stevens, Andrea J. Pruijssers, et al., Durability of responses after sars-cov-2 mrna-1273 vaccination, N. Engl. J. Med. 384 (1) (2021) 80–82.

[7] Merryn Voysey, Sue Ann Costa Clemens, Shabir A. Madhi, Lily Y. Weckx, Pedro M. Folegatti, Parvinder K. Aley, Brian Angus, Vicky L. Baillie, Shaun L. Barnabas, Qasim E. Bhorat, et al., Safety and efficacy of the chadox1 ncov-19 vaccine (azd1222) against sars-cov-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK, Lancet 397 (10269) (2021) 99–111.

[8] Christian Gaebler, Zijun Wang, CC Lorenzi Julio, Frauke Muecksch, Shlomo Finkin, Minami Tokuyama, Alice Cho, Mila Jankovic, Dennis Schaefer-Babajew, Thiago Y. Oliveira, et al., Evolution of antibody immunity to sars-cov-2, Nature (Lond.) 591 (7851) (2021) 639–644.

[9] Shuhui Song, Lina Ma, Zou Dong, Dongmei Tian, Cuiping Li, Junwei Zhu, Meili Chen, Anke Wang, Yingke Ma, Mengwei Li, et al., The global landscape of sars-cov-2 genomes, variants, and haplotypes in 2019ncovr, Dev. Reprod. Biol. 18 (6) (2020) 749–759.

[10] O.G. Pybus, et al., O'Toole Á, Hill V. Tracking the international spread of sars-cov-2 lineages b.1.1.7 and b.1.351/501y-v2 [version 1; peer review: 3 approved], Wellcome Open Res 6 (2021) 121.

[11] Adam S. Lauring, Emma B. Hodcroft, Genetic variants of sars-cov- 2—what do they mean? JAMA 325 (6) (2021) 529–531.

[12] Carolina KV. Nonaka, Marília Miranda Franco, Tiago Gräf, Camila Araújo de Lorenzo Barcia, Renata Naves de Ávila Mendonça, Karoline Almeida Felix De Sousa, Leila Mc Neiva, Vagner Fosenca, Ana VA. Mendes, Santana Renato De Aguiar, et al., Genomic evidence of sars-cov-2 reinfection involving e484k spike mutation, Brazil, Emerg. Infect. Dis. 27 (5) (2021) 1522.

[13] Allison J. Greaney, Tyler N. Starr, Pavlo Gilchuk, Seth J. Zost, Elad Binshtein, Andrea N. Loes, Sarah K. Hilton, John Huddleston, Rachel Eguia, Katharine HD. Crawford, et al., Complete mapping of mutations to the sars-cov-2 spike receptor-binding domain that escape antibody recognition, Cell Host Microbe 29 (1) (2021) 44–57.

[14] Emma C. Thomson, Laura E. Rosen, James G. Shepherd, Roberto Spreafico, Ana da Silva Filipe, Jason A. Wojcechowskyj, Chris Davis, Luca Piccoli, David J. Pascall, Josh Dillen, et al., Circulating sars-cov-2 spike n439k variants maintain fitness while evading antibody-mediated immunity, Cell 184 (5) (2021) 1171–1187.

[15] Carl A. Ascoli, Could mutations of sars-cov-2 suppress diagnostic detection? Nat. Biotechnol. 39 (3) (2021) 274–275.

[16] Bai Chen, Arieh Warshel, Critical differences between the binding features of the spike proteins of sars-cov-2 and sars-cov, J. Phys. Chem. B 124 (28) (2020) 5907–5912.

[17] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine HD. Crawford, S Dingens Adam, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, et al., Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding, Cell 182 (5) (2020) 1295–1310.

[18] Houriiyah Tegally, Eduan Wilkinson, Marta Giovanetti, Arash Iran-zadeh, Vagner Fonseca, Jennifer Giandhari, Deelan Doolabh, Suresh-nee Pillay, Emmanuel James San, Nokukhanya Msomi, et al., Detection of a sars-cov-2 variant of concern in South Africa, Nature (Lond.) 592 (7854) (2021) 438–443.

[19] Petra Mlcochova, Steven Kemp, Mahesh Shanker Dhar, Papa Guido, Bo Meng, Swapnil Mishra, Charlie Whittaker, Thomas Mellan, Isabella Ferreira, Rawlings Datir, et al., Sars-cov-2 B. 1.617. 2 Delta Variant Emergence and Vaccine Breakthrough, 2021.

[20] Maxwell Salvatore, Rupam Bhattacharyya, Soumik Purkayastha, Lauren Zimmermann, Debashree Ray, Aditi Hazra, Michael Kleinsasser, Thomas Mellan, Charlie Whittaker, Seth Flaxman, et al., Resurgence of Sars-Cov-2 in india: Potential Role of the B. 1.617. 2 (Delta) Variant and Delayed Interventions, MedRxiv, 2021.

[21] Wenyang Zhou, Chang Xu, Meng Luo, Pingping Wang, Zhaochun Xu, Guangfu Xue, Xiyun Jin, Yan Huang, Yiqun Li, Huan Nie, et al., Mutcov: a pipeline for evaluating the effect of mutations in spike protein on infectivity and antigenicity of sars-cov-2, Comput. Biol. Med. 145 (2022), 105509.

[22] Puneet Rawat, Divya Sharma, Medha Pandey, R. Prabakaran, M Michael Gromiha, Understanding the mutational frequency in sars-cov-2 proteome using structural features, Comput. Biol. Med. (2022), 105708.

[23] Baishali Mullick, Rishikesh Magar, Aastha Jhunjhunwala, Amir Barati Farimani, Understanding mutation hotspots for the sars- cov-2 spike protein using shannon entropy and k-means clustering, Comput. Biol. Med. 138 (2021), 104915.

[24] Abdullah Shah, Saira Rehmat, Iqra Aslam, Muhmmad Suleman, Farah Batool, Abdul Aziz, Farooq Rashid, Muhmmad Asif Nawaz, Syed Shu- jait Ali, Muhammad Junaid, et al., Comparative mutational analysis of sars-cov-2 isolates from Pakistan and structural-functional implications using computational modelling and simulation approaches, Comput. Biol. Med. 141 (2022), 105170.

[25] Zhen Xia, Gulei Jin, Jun Zhu, Ruhong Zhou, Using a mutual information-based site transition network to map the genetic evolution of influenza a/h3n2 virus, Bioinformatics 25 (18) (2009) 2309–2317.

[26] Bai Chen, Junlin Wang, Geng Chen, Honghui Zhang, An Ke, Peiyi Xu, Du Yang, Richard D. Ye, Arjun Saha, Aoxuan Zhang, et al., Predicting mutational effects on receptor binding of the spike protein of sars-cov-2 variants, J. Am. Chem. Soc. 143 (42) (2021) 17646–17654.

[27] Rui Yin, Emil Luusua, Dabrowski Jan, Yu Zhang, Chee Keong Kwoh, Tempel: time-series mutation prediction of influenza a viruses via attention-based recurrent neural networks, Bioinformatics 36 (9) (2020) 2697–2704.

[28] M Cyrus Maher, Istvan Bartha, Steven Weaver, Julia Di Iulio, Elena Ferri, Leah Soriaga, Florian A. Lempp, Brian L. Hie, Bryan Bryson, Bonnie Berger, et al., Predicting the mutational drivers of future sars- cov-2 variants of concern, Sci. Transl. Med. 14 (633) (2022), eabk3445.

[29] Brian Hie, Ellen D. Zhong, Bonnie Berger, Bryan Bryson, Learning the language of viral evolution and escape, Science 371 (6526) (2021) 284–288.

[30] Ehsaneddin Asgari, Mohammad RK. Mofrad, Continuous distributed representation of biological sequences for deep proteomics and genomics, PLoS One 10 (11) (2015), e0141287.

[31] Yuelong Shu, John McCauley, Gisaid: global initiative on sharing all influenza data–from vision to reality, Euro Surveill. 22 (13) (2017), 30494.

[32] Matthew McCallum, Anna De Marco, Florian A. Lempp, M. Alejandra Tortorici, Dora Pinto, Alexandra C. Walls, Martina Beltramello, Alex Chen, Zhuoming Liu, Fabrizia Zatta, et al., N-terminal domain antigenic mapping reveals a site of vulnerability for sars-cov-2, Cell 184 (9) (2021) 2332–2347.

[33] Kazutaka Katoh, Daron M. Standley, Mafft multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (4) (2013) 772–780.

[34] Corinna Cortes, Vladimir Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[35] David W. Hosmer Jr., Lemeshow Stanley, Rodney X. Sturdivant, Applied Logistic Regression, ume 398, John Wiley & Sons, 2013.

[36] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[37] Wojciech Zaremba, Ilya Sutskever, Oriol Vinyals, Recurrent Neural Network Regularization, 2014 *arXiv preprint arXiv:1409.2329*.

[38] Sepp Hochreiter, Jürgen Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[39] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Thirion Bertrand, Olivier Grisel, Mathieu Blondel, Prettenhofer Peter, Ron Weiss, Dubourg Vincent, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[40] Paszke Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, Lerer Adam, Automatic Differentiation in Pytorch, 2017.

[41] Yiming Bao, Pavel Bolotov, Dmitry Dernovoy, Boris Kiryutin, Leonid Zaslavsky, Tatiana Tatusova, Jim Ostell, David Lipman, The influenza virus resource at the national center for biotechnology information, J. Virol. 82 (2) (2008) 596–601.