ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

Research on An Antigen Variation Prediction Model Based on the IPA-VGG Model

Rongting Li[1, a], Xiaozhou Chen[1, b, *], Rui Hu[1, c], Quanfeng Xu[1, d], Peng Duan[1, e]

[1]Yunnan Minzu University, Mathematics and Computer Science, Kunming, China

[a]ai_blade@163.com, [b]ch_xiaozhou@163.com, [c]harry919@126.com, [d]xqf3520@163.com, [e]1654964109@qq.com

Abstract

The rapid evolution of influenza viruses continuously leads to the emergence of new influenza strains, which can escape the human immune surveillance system, making thus the on-time detection of antigen variation a key for the design of vaccines. Traditional methods, such as the haemagglutination inhibition (HI) assays, are time-consuming, labor-intensive and require the availability of a mutated live virus. With the advancements in computer and information technology, many computational models have been developed to predict antigen variation. In addition, influenza sequences distributed by similar residues present high similarity, which may affect the prediction results. Therefore, a two – dimensional (2D) convolutional network model, called IPA-VGG, was introduced to infer influenza antigen variation considering the above-mentioned fact. In particular, IPA(Improved ProVect of Apriori), a new amino acid encoding method, is introduced to multiple downstream proteomes to facilitate machine learning analysis. After analyzing the sequence of influenza virus strains, a two-dimensional convolution-excitation VGG structure was constructed, which fused the virus sequence and encoded information to focus the network on information residual features. Experimental results on influenza A (H1N1) datasets demonstrated that the IPA-VGG model showed superior performance when combined with the new encoding method and the convolutional architecture. When compared with a traditional machine learning model, the convolutional neural network model based on IPA amino acid coding presented better performance. Therefore, it is inferred that this model can be used as a reliable and robust tool for predicting antigen variation.

Keywords

VGG; Antigenic variation; Deep learning; Apriori.

1. INTRODUCTION

Seasonal influenza poses a serious threat to public health and the global economy, killing up to half a million people and sickening millions more each year worldwide. H1N1 and H3N2 are the two main subtypes of influenza viruses that infect human populations, while vaccination is the most effective way to prevent infection with influenza viruses. However, the components of the vaccine must be updated regularly to ensure its effectiveness. Influenza virus surface glycoprotein hemagglutinin (HA) is the main target of host immune [1]. However, the accumulation of HA protein mutations leads to the emergence of new antigenic variants that cannot be effectively suppressed by antibodies, posing a huge challenge for the design of vaccines. The development of rapid and reliable influenza antigenicity assays is the key to influenza vaccine design and influenza surveillance.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

Until now, about 180 kinds of amino acids have been discovered, but only 20 of them are commonly used in the synthesis of peptide chains, see Table 1 [2]. The aforementioned HA sequence is a protein sequence that carries the antigenic properties of the virus [3]. The current research is mainly emphasizing the prediction of antigen mutations.

Table 1. Basic amino acids and their abbreviations

Name	abbreviations	Name	abbreviations
alanine	A	histidine	H
leucine	L	cysteine	C
phenylalanine	F	serine	S
Complex amino acid	Y	glutamine	Q
asparagine	N	glutamate	E
lysine	K	aspartic acid	D
methionine	M	tryptophan	W
threonine	T	B leucine	I
proline	P	valine	V
arginine	R	glycine	G

Deep neural networks have already been successfully applied in many fields, including bioinformatics. Convolutional Neural Network (CNN) is one of the most commonly used machine learning methods in bioinformatics problems, including classification of efflux proteins from membrane [4], human leukocyte antigen class I-peptide binding prediction [5], prediction of protein secondary structure [6] and prediction of protein-protein interaction [7]. In the paper, we use deep learning techniques, initially designed for natural language processing (NLP), to solve the problem of influenza A virus antigen variation prediction. In particular, two different antigen sequences are compared, according to the antigen and its classification. Then based on the work of Rui [8], we further optimized the encoding via Provect method, introducing a new encoding amino acid, named IPA, and mapping it to a triple amino acids-100 dimensional vector space. Considering that most neurons are in the inhibited state during the neural network operation, this paper suggests a method, called IPA-VGG, which combines the two-dimensional CNN model with the convolution-activation mechanism, for the prediction of antigen variation.

2. RELATED TECHNOLOGIES

2.1. Hemagglutinin (HA)

Influenza vaccine is mainly HA-based. HA, encoded by the gene segment number 4, is the most important protein of all influenza viruses. Its main function is to bind to the host cell receptor and to penetrate the cell membrane of the host allowing it to escape the surveillance of the host immune system.

2.2. Antigenic Distance

Antigenic distance is a quantitative metric measuring whether and how extensively an antigen has been mutated. The antigenic distance between the two strains was a calculation method that was initially suggested by Archetti-Horsfall [9], and it can be calculated using the following formula, see Eq. (1):

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

$$D_{ij} = \sqrt{\frac{H_{ii} \times H_{jj}}{H_{ij} \times H_{ji}}} \quad (1)$$

H_{ij} being the maximum dilution factor of the cell agglutination induced by the antiserum of strain i to strain j. Liao et al. [10] defined 4 as the threshold value. If it is equal to or greater than 4, strain i and strain j are defined as having different antigens. Otherwise, the pair of strains are considered antigenically similar. The median titer value was used to calculate the antigen distance for duplicate pairs of strains whose HI titers were measured at multiple independent institutions. 937 pairs of different antigens and 257 pairs of similar antigens were found in the A (H1N1) virus sequence. In this study, we ended up with 294 unique H1N1 sequences. The amino acid numbers of these different subtypes of protein sequences were recommended by Burke and Smith [11].

2.3. K-mer

mer is a coding method that equivalent represents a gene sequence by calculating the frequencies of K-linked nucleotides [12]. K-mer is a very effective coding method. Almost all coding methods are extended on the basis of K-mer. On this basis, this paper chooses three amino acids to represent the equivalent protein sequence fragments, and assumes that our amino acid sequence is DTLCIGYHANNSTD, with a total length of 14bp. In this protein sequence, DTL, TLC, LCI, CIG, IgY, GyH, YHA, Han, Ann, NNS, NST, and STD were included as triamino acid units. The frequency of any triplet amino acid can be calculated as shown in Equation, see Eq. (2):

$$P_{abc} = \frac{N_{abc}}{l-k+1} \quad (2)$$

Where Pabc represents the occurrence frequency of triple amino acid abc, a, b and c are any amino acids in Table 1, Nabc represents the frequency of occurrence of the triple amino acid abc, l is the total length of nucleotides, and k is 3. Thus, a specific protein sequence can be transformed into a feature vector to represent it.

2.4. Convolutional Neural Network

Convolutional Neural Network (CNN) is a subtype of Feedforward Neural Networks with deep structure topology. CNN contains convolutional computation and is one of the representative algorithms of deep learning. CNNs have been widely used in image processing [13], but the superior self-learning ability can also be applied to text data. The present research work optimizes the existing convolutional neural network and reasonably modifies the input hidden and output layers to allow the CNN topology to meet the prediction requirements and purpose.

2.5. Headings

Level one headings for sections: should be in bold, flushed to the left, and numbered using Arabic numbers, such as 1, 2, ….

Level two headings for subsections: should be in bold-italic, flushed to the left, and numbered after the level one heading. For example, the second level two heading under the third level one heading should be numbered as 3.2.

Level three headings: should be in italic, flushed to the left, and similarly, numbered after the level two headings, such as 3.2.1, 3.2.2, etc.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

The Initial letter of each notional word in all headings is capitalized.

3. PREDICTION ALGORITHM BASED ON IPA CODING MANNER AND VGGNET

3.1. IPA Coding Manner

In the ProVect coding method [14], most of the trivalent amino acids can find the corresponding 100-dimension feature vectors and re-encode them, and import the feature vectors of the corresponding sites of two different antigen sequences into the CNN by difference, while selectively ignoring the special amino acids that cannot be recognized in a few antigen sequences. Based on the above situation, the Improved ProVect of Apriori (IPA) pretreated the dataset in this experiment for the special amino acids "B", "J", "Z" and "X", that could not be recognized in the sequence of antigen using a substitution matrix based on hydrophilicity proposed by Meng [2], to identify the substitutive amino acid fragments containing the above four amino acids.

HA of influenza virus has high variability and there are mainly four antigenic determinants, namely Ca (137-142,166-170,203,221,235), CB (70-75), SA (124,153-157,159-164) and Sb (184-195). These 44 amino acids are the target of the host immune system, and their variation affects the antigenic drift of the virus [15]. A support count is proposed to measure the relationship between the above-mentioned commonly used and the currently used amino acids using the following formula, see Eq. (3):

$$\sigma(x) = \left|\left\{t_i \middle| x \subseteq t_i, t_i \in T\right\}\right| \tag{3}$$

Where the symbol represents the number of elements in the set, T is the set of all antigenic Strain alignment sequences ti, and x is the item set of all comparison sequences containing a particular amino acid in the four antigenic determinants. Based on the above formula, the minimum support count was calculated to be 30. Combined with the Apriori algorithm [16] and taking the unusual amino acid B as an example, the associated amino acid of B can be obtained, as depicted in Figure 1.

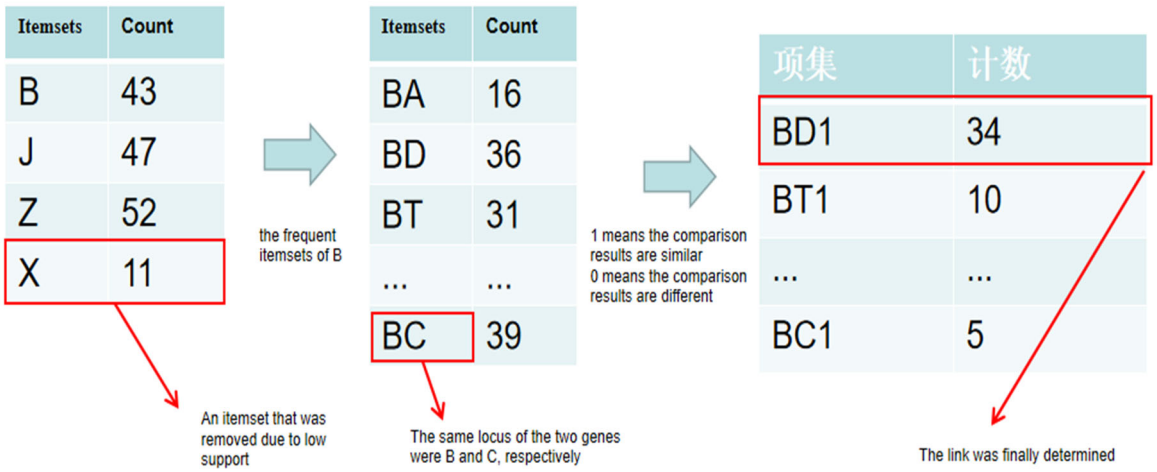

Figure 1. Schematic diagram of the Apriori algorithm

According to the Apriori algorithm, it can be inferred that the amino acid sites B, J and Z can be replaced by Nd, IL and QE, respectively. X is randomly replaced by the common amino acids

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

in the following experiment due to its small size and insignificant correlation degree. This paper improved the IPA after encoding used for each of the joint recoding amino acids for 100 d characteristic vector, as much as possible in order to restore the data features with the new encoding being composed of two different antigen sequences of the same site feature vector for data integration. In contrast to the research of in Liao [10] the aim of the current manuscript's approach was to retain as much as possible of the characteristics of the two different antigen sequences, as shown in the flow chart of Figure 2.

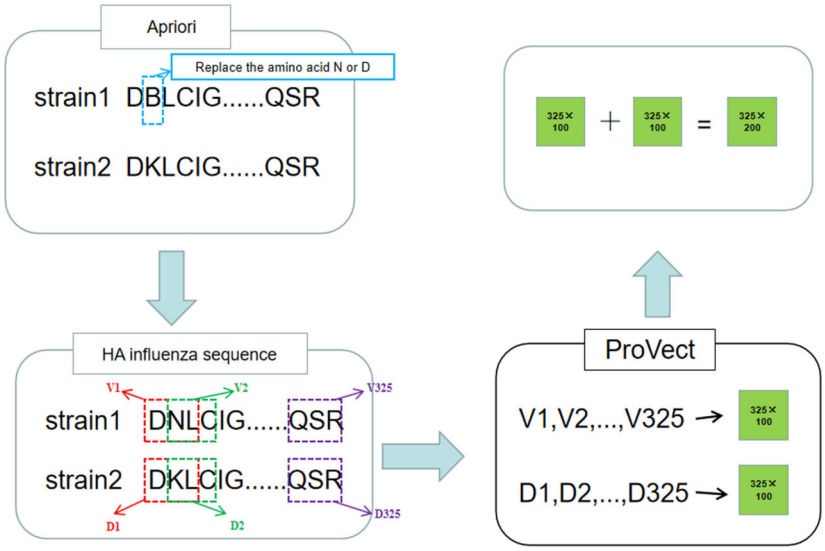

Figure 2. Flow chart of IPA coding method

3.2. Training Models of VGGNet

At present, convolutional neural networks have been widely used in image processing [17]. However, their superior algorithm and flexible structure have great potential and plasticity in the field of deep learning allowing them to be applied in other domains. Based on this technical background, CNN method was selected for prediction, and on this basis, VGG13 network optimization was used to achieve the purpose of reducing the dimension of text data, finding its features through the internal connection of data and performing classification.

The two antigen sequences were recoded and integrated according to the above-mentioned coding method and a 325 × 200 matrix was obtained. This matrix is transformed into a square matrix by using the 2D convolutional neural network function encapsulated by TensorFlow in Python and then connected with other hidden layers. The first convolution-pooling contains 64 convolution kernels, the second convolution-pooling contains 128 convolution kernels, the third convolution-pooling contains 256 convolution kernels, the fourth convolution-pooling contains 512 convolution kernels and the fifth convolution-pooling contains 512 convolution kernels with each convolution kernels having 3 × 3 size to ensure that each convolution-pooling is small and deep and to guarantee high accuracy [18]. A forgetting layer is added after each convolution-pooling to carry out random forgetting with a forgetting rate of 0.25 to ensure that the model will not be overfitted on the training data. Subsequently, a layer of laminating layer is added before the full connection layer to convert the multi-dimensional input into a one-dimensional one [19]. Finally, the classification layer is connected. In summary, the size diagram and flow chart between the above-mentioned layers are shown in Figure 3 and Figure 4.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

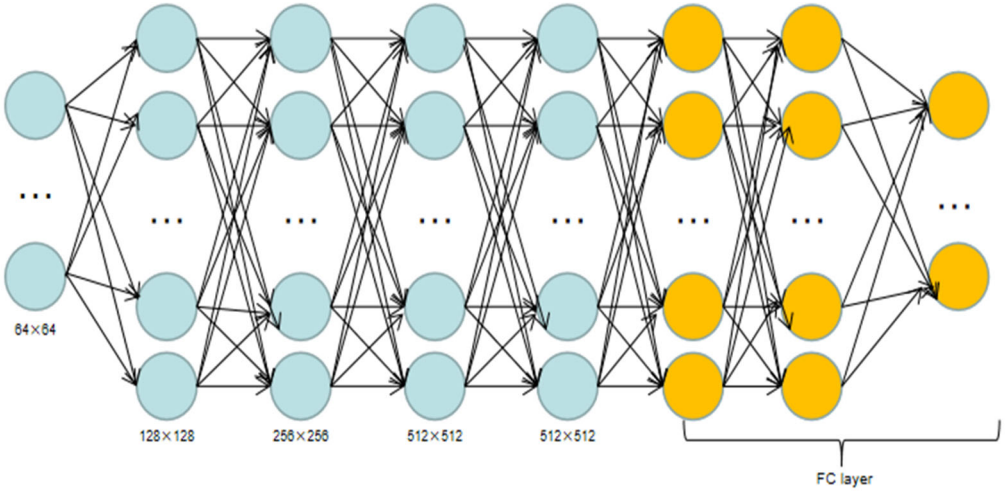

Figure 3. Schematic diagram of network size

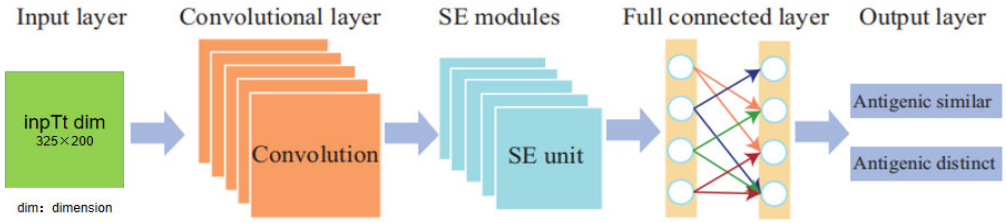

Figure 4. IPA-VGG work flow chart

The multiple addition of forgetting layer inhibits the occurrence of overfitting in comparison with the traditional convolutional neural network. The previous work maintained a difference between the eigenvectors, but in the implementation of this manuscript the eigenvectors of the three amino acids are arranged horizontally in the same position, which is bound to greatly increase the workload of training. Therefore, all activation functions in this model are selected as ReLU functions, which greatly increases the computational resources required for training on the premise of increasing the accuracy. that the selection of SGD [20] as the optimizer (decay=1e-6, Momentum =0.9, Nesterov =True) for the data was found to provide the most accurate predictions. The detailed steps of the rough IPA-VGG model are shown in Table 2:

4. EXPERIMENT AND ANALYSIS

The experimental data were compared with the A (H1N1) virus dataset and the A (H1N1) antigen HA sequence dataset published on GitHub. A computer with 8G memory and AMD Ryzen 5 CPU (main frequency 3.4 GHz) was selected as the hardware platform for the experiment and the Google cloud server was used for the analysis under Windows 10, 64-bit operating system.

4.1. Data Set Introduction

The datasets used in this article were retrieved from the publicly available repository GitHub. There were 1563 datasets of original antigen Strain alignment (for each pair, the result is Distance), and 294 datasets of the HA sequence sequences for the corresponding antigen Strain (denoted as one for each antigen).

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

Table 2. Detailed steps of IPA-CNN model for predicting antigen variation

Require:A pair of influenza HA1 sequence a and b

Ensure: Antigenic relationship between a and b

Feature generation (Section 2.3)
n←The length of HA1 protein
for i =1 to n do
  a_i, b_i←Splittings for strains a and b
  IPA(a_i),IPA(b_i)←Embedding vectors for subsequences ai and bi
  v_i =[IPA(a_i)] +[IPA(b_i)]←The sum vector of two subsequences ai and bi
end for
V =[v_1, ..,v_n]←The representation of two stains a and b
X, Y←The training samples through feature space V
for i= 1 to epoch do
  Do initializationnet
  Net = train (IPA-VGG, parameters)
  for j =1 to numbatches do
    Xbatch =X [j: j+ batchsize,:,:,:]
    Ybatch =Y [j: j+ batchsize]
    scores = net (X_{batch})
    loss = Cross Entropy Loss (scores, Ybatch)
    Optimizer.step()
    Predictions =Output(scores)
  end for
end for
return Predictions

4.2. Data Preprocessing

The original data were processed as follows: (1) Classification was performed according to the distance on the Strain comparison data of different antigens; (2) Data cleaning was conducted to remove null values; (3) The antigen Strain was translated to its corresponding HA amino acid sequence; (4) Re-encoding was applied using the IPA encoding method; (5) The dataset was divided into training set, verification set and test set in an 8:1:1 ratio according to the partition basis of Tan [12].

4.3. Experimental Results

In order to verify the effectiveness of the proposed IPA encoding, The present article, applied support vector machine (SVM), random forests (RF), logistic regression (LR), K nearest neighbor (KNN) and neural network (NN), as commonly used classification models, after preprocessing the raw data, to verify the classification results. Experimental results showed that the coding method can effectively improve the classification accuracy. Considering the unbalanced nature of the dataset, five parameters were used as the evaluation criteria of the different applied methods, including accuracy, recall rate, F-score [21] and Matthews Correlation Parameter (MCC) [22]. According to the measure ξ, the proposed model is superior, and the relevant formula for this metric is as follows, see Eq. (4).

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

$$\xi = \begin{cases} 0 & ,MCC \le 0 \\ F1 \times (\text{rec} + \text{acc} + \text{pre}), & MCC > 0 \end{cases} \tag{4}$$

At the same time, the memory requirements comparison between the above results and the results of the CNN model proposed in this paper confirmed that the newly introduced model was proved to be better than the previous methods on this basis. The proposed model's results are shown in Figure 5.

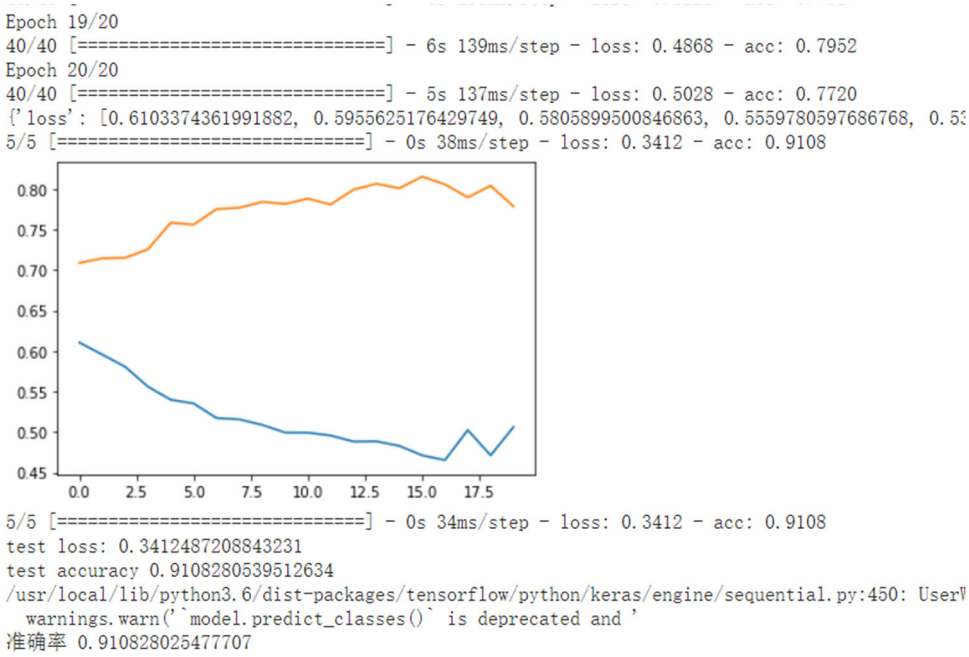

Figure 5. Results Display Diagram

And the comparative prediction results are shown in Table 3.

Table 3. Comparison of experimental results

Model	Training data					Testing data					
	Accuracy	Precision	recall	F-score	MCC	Accuracy	Precision	recall	F-score	MCC	ξ
SVM	0.594	0.594	1.000	0.745	0.409	0.623	0.623	1.000	0.768	0.423	1.725
RF	0.987	0.993	0.985	0.989	0.974	0.863	0.884	0.897	0.891	0.706	2.356
LR	0.817	0.816	0.892	0.853	0.616	0.722	0.752	0.826	0.787	0.392	1.810
KNN	0.901	0.956	0.873	0.853	0.616	0.722	0.752	0.826	0.787	0.392	1.810
NN	0.998	0.997	0.999	0.998	0.995	0.859	0.895	0.877	0.886	0.703	2.331
IPA-CNN	0.911	0.916	0.985	0.949	0.596	0.911	0.918	0.985	0.931	0.526	2.619

5. CONCLUSION

Based on the comparative data of a variety of A (H1N1) antigen HA, the present manuscript carried out a study on the prediction of influenza antigen variation. The internal structure of CNN model was improved by introducing IPA encoding method, the, which effectively improved the prediction accuracy in the problem of predicting whether influenza A (H1N1) virus mutation occurred, raising the creditability of the overall approach.

The following improvements were made based on the realization of different A (H1N1) antigens and their sequence data under IPA encoding method: (1) a further optimized IPA

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

encoding method was introduced by using an alternative confidence function combined with the Provect encoding method; (2) the abnormal amino acids was substituted for common amino acids according to the hydrophobic matrix of amino acids; (3) IPA coding method was combined with traditional machine learning model to verify the validity; (4) A deep convolutional neural network model was built using the newly encoded data.

Further research can be done on this model: (1) Influenza virus is not only a subtype of influenza A H1N1 and thus the prediction analysis should be conducted for other subtypes of influenza viruses; (2) The method of dealing with abnormal amino acids in this model is only applicable to this dataset, and it is difficult to guarantee that the influenza virus will not mutate again in the future. Therefore, this simulation model has timeliness and should be improved to include prognostic prediction analysis; (3) CNN model possesses among others the characteristics of high plasticity and fast updating speed. On this basis, the optimization process of CNN model is bound to be a difficult and time consuming, making the improvement of the model itself a necessity.

ACKNOWLEDGMENTS

This paper was supported by the National Natural Science Foundation of China (Grant No. 31460297). The authors would like to thank the editor in chief and worthy referees for valuable suggestions for giving the final shape of the manuscript.

REFERENCES

[1] R. Y. Gao, M. Gu, L. W. Shi, K. T. Liu, X. L. Li, X. Q. Wang, J. Hu, X. W. Liu, S. L. Hu, S. J. Chen, D. X. Peng, X. N. Jiao and X. F. Liu, N-linked glycosylation at site 158 of the HA protein of H5N6 highly pathogenic avian influenza virus is important for viral biological properties and host immune responses, Veterinary ResearchVolume, vol. 52, no. 1, pp. 2-14, 2021.

[2] X. Y. Meng, J. Meng and J. L. Ge, An alternative matrix based on hydrophilicity and hydrophobicity, MATHEMATICS IN PRACTICE AND THEORY, vol.39, No.7, 2009.

[3] X. W. Ren, Y. F. Li, X. N. Liu, X. P. Shen, W. L. Gao and J. S. Li, Computational identification of antigenicity-associated sites in the hemagglutinin protein of a/h1n1 seasonal influenza virus, PloS one, vol. 10, no. 1, 2015.

[4] S. W. Taju, T. Nguyen, N. Le, R. M. I. Kusuma and Y. Ou,Deepefflux: a 2d convolutional neural network model for identifying families of efflux proteins in transporters, Bioinformatics, vol. 34, no. 18, p. 3111–3117, 2018.

[5] Y. S. Vang and X. H. Xie, Hla class i binding prediction via convolutional neural networks, Bioinformatics, vol. 33, no. 17, p. 2658–2665, 2017.

[6] Z. Li and Y. Z. Yu, Protein secondary structure prediction using cascaded convolutional and recurrent neural networks, International Joint Conference on Artificial Intelligence (IJCAI), 2016.

[7] T. Sun, B. Zhou, L. H. Lai and J. F. Pei, Sequence-based prediction of protein interaction using a deep-learning algorithm, BMC bioinformatics, vol. 18, no. 1, p. 277, 2017.

[8] R. Yin, X. R. Zhou, F. X. Ivan, J. Zheng, V. T. K. Chow and C. K. Kwoh, Identification of potential critical virulent sites based on hemagglutinin of influenza a virus in past pandemic strains, the 6th International Conference, 2017.

[9] W. Ndifon, J. Dushoff and S. A. Levin,On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness, Vaccine, vol. 27, no. 18, pp. 2447–2452, 2009.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0025

[10] Y. C. Liao, M. S. Lee, C. Y. Ko and C. A. Hsiung. "Bioinformatics models for predicting antigenic variants of influenza a/h3n2 virus," Bioinformatics, vol. 24, no. 4, p. 505–512, 2008.

[11] D. F. Burke and D. J. Smith, A recommended numbering scheme for influenza a ha subtypes,"PloS one, vol. 9, no. 11, 2014.

[12] Z. Y. Tan, Predicting the Antigenic Variant of Influenza Virus and Virus Host Based on Deep Learning Methods, M.S. dissertation, University of Hunan, China.

[13] J. Wang, Y. Yang, J. H. Mao, Z. H. Huang, C. Huang and W. Xu, Cnn-rnn: A unified framework for multi-label image classification, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 2285–2294, 2016.

[14] E. Asgari, M. RK. Mofrad, Continuous distributedrepresentation of biological sequences for deep proteomics and genomics, PloS one, vol. 10, no. 11, 2015.

[15] X. Zhao, Z. Teng, F. H. Fang, F. Yuan, C. Y. Jiang, Z. G. Yuan and X. Zhang, Antigenic and genetic characteristics of influenza B virus/Victoria lineage in Shanghai 2018–2020, DISEASE SURVEILLANCE, vol. 35, no. 12, p. 1074-1080, 2020.

[16] P. H. Lu, J. L. Keng and K. Kuo, An Apriori Algorithm-Based Association Rule Analysis to Identify Herb Combinations for Treating Uremic Pruritus Using Chinese Herbal Bath Therapy, Evidence-based complementary and alternative medicine, vol. 2020, no. 7, p. 1-9, 2020.

[17] A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, In Advances in neural information processing systems, p. 1097–1105, 2012.

[18] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, Wei Xu, Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, p. 2285–2294, 2016.

[19] K. Simonyan and A. Zisserman, Very deep convolutional networks for arge-scale image recognition, Computer Science, 2014.

[20] L. Bottou, Large-scale machine learning with stochastic gradient descent, Physica-Verlag HD , Springer, p. 177–186, 2010.

[21] C. Goutte and E. Gaussier, A probabilistic interpretation of precision, recall and f-score, with implication for evaluation, In European Conference on Information Retrieval, Springer, p. 345–359, 2005.

[22] M. Bekkar, H. K. Djemaa and T. A. Alitouche, Evaluation measures for models assessment over imbalanced data sets, Journal of Information Engineering and Applications, vol. 3, no. 10, 2013.