https://www.scirp.org/journal/jbise

J. Biomedical Science and Engineering, 2020, Vol. 13, (No. 8), pp: 175-186

Prediction of Mutations in H7 Hemagglutinins from Influenza A Virus

Shaomin Yan, Guang Wu

National Engineering Research Center for Non-Food Biorefinery, State Key Laboratory of Non-Food Biomass and Enzyme Technology, Guangxi Biomass Engineering Technology Research Center, Guangxi Key Laboratory of Bio-Refinery, Nanning, China

Correspondence to: Guang Wu, hongguanglishibahao@gxas.cn

Keywords: Hemagglutinin, Influenza, Mutation, Neural Network, Prediction, Randomness

Received: June 19, 2020 **Accepted:** August 10, 2020 **Published:** August 13, 2020

ABSTRACT

Influenza A viruses have led several pandemics and epidemics in human history. H7 subtype influenza mainly infects avian but also humans occasionally. Since the outbreak of H7N9 subtype influenza occurred in China in 2013, this virus is still circulating in domestic poultry and leading several waves of influenza. To prevent influenza, vaccination is an important strategy. However, influenza virus evolves constantly, but unpredictably. If we would have a one-to-one cause-mutation relationship, the mutation prediction would be possible. However, many external causes, which led to the mutations in the past, might not leave any trace due to the change in environments, whereas the current virus might not be subject to the historically external causes because of evolution. Furthermore, the protein should have the internal causes, which might be quite unclear and difficult to quantify, to engineer mutations. Indeed, various forces twist proteins into 3-demensional structures, whereas any perturbation could lead to a mutation. Of various internal causes for mutation, randomness in protein primary structure should play an important role in mutation. Over years, we have developed three methods to quantify the randomness within a protein primary structure; thus we build a relationship between cause, which is randomness in primary structure, and mutations, which are occurrence and non-occurrence of mutation. In this way, the cause-mutation relationship becomes the problem of classification, which can be solved using logistic regression and neural network. In this study, we apply this model to predict 1) the mutation positions in H7 hemagglutinins from influenza A virus and 2) the would-be-mutated amino-acids at predicted positions with the amino-acid mutating probability. The results show suitability and predictability in such modelling, and pave the way for further development.

1. INTRODUCTION

Influenza viruses have led several pandemics and epidemics in human history [1, 2]. Of various subtypes of influenza A virus, H7 subtype mainly infects avians but also humans occasionally [3]. The first outbreak of H7 subtype influenza simultaneously infecting poultry and humans occurred in 2003 [4]. In 2013, H7N9 subtype influenza occurred in China [5, 6], and since then H7N9 subtypes influenza virus has been circulating in domestic poultry and leading several waves of influenza in China [7, 8].

To prevent these constant, but different sized epidemics, vaccination is an important strategy [9]. However, influenza virus evolves constantly, but unpredictably [10]. If the mutations in influenza virus would be predictable, then vaccination would be more applicable.

Certainly, the best way to predict the mutation is to find the cause for mutation, and then we can build a one-to-one cause-mutation relationship. Thereafter, we can predict the mutation if the same cause appears again.

However, this approach might not work well, because 1) many causes, which led mutations in the past, might not leave any cue to us due to the huge changes in environments; 2) the conditions, under which the historical causes functioned, might never be known due to the fact that the conditions, which are defined by modern technique, might be impossible to be determined by the technique in the past; and 3) the current virus might not be subject to the historically external causes, which led the mutations in the past, because of evolution.

Furthermore, we might consider that there are internal causes within virus, for example, viral proteins, because various forces twist proteins into 3-demensional structures, whereas any perturbation could lead to a mutation. Of various internal causes for mutation, randomness in protein primary structure should play an important role in mutation because pure chance is now considered to lie at the very heart of nature [11]. We could establish a cause-mutation relationship to predict the mutations engineered by internal randomness if we could define and quantify the randomness within DNA/RNA/protein. Over years, we have developed three methods to quantify the randomness within a protein primary structure [12-15]; thus we build a relationship between cause, which is randomness in primary structure, and mutations, which are occurrence and non-occurrence of mutation. In this way, the cause-mutation relationship becomes the problem of classification, which can be solved using logistic regression and neural network.

With our quantified randomness, we could predict mutations using the cause-mutation relationship, because we can classify the occurrence and non-occurrence of mutations as unity and zero. Thus we switch the cause-mutation relationship to the classification problem, which can be solved using either statistical tool such as logistic regression [16, 17] or neural network tool such as perceptron and backpropagation [18].

The hemagglutinin is the major surface antigen of influenza viruses, against which neutralizing antibodies are elicited during virus infection and vaccination [19]. Of 15 subtypes of hemagglutinins, H7 hemaglutnin has its special representative in structure [20]. The domestic pigeons are particularly subject to H7 hemagglutinin, and infected pigeons can act as mechanical vectors for long-distance transmission [21]. Moreover, it is now considered that H7 contributes the amantadine resistance [22].

In fact, the prediction of mutation should include at least two steps; say, the prediction of mutation position and the prediction of would-be-mutated amino acids at predicted positions. In this study, we try to predict 1) the mutation positions in H7 hemagglutinins from influenza A virus and 2) the would-be-mutated amino-acids at predicted positions with the amino-acid mutating probability.

2. MATERIALS AND METHODS

Data and Their Elaboration

170 H7 hemagglutinins from influenza A virus were obtained from the influenza virus resources [23].

The data were grouped according to their sampling places and time for initial inspection. Furthermore, an evolutional relationship was found by means of phylogenetic analysis [24]. Along the phylogenetic tree, the H7 hemagglutinin sequences were clustered into parental relationships, *i.e.* father and

daughter. Such parental relationship marks the difference in H7 hemagglutinin sequences, *i.e.* mutations between father and daughter sequences.

Then the data were divided into two groups [25], one group worked as training group with more advanced years before 2000 while the other group worked for the prediction and validation with years after 2000.

Prediction model

Ideally, we hope that the prediction model would function in such a way, that is, we input a protein sequence into the model, whose output is the predicted mutation position. Naturally we should use the historical data to train the prediction model.

As we have developed three methods to quantify the randomness within a protein, thus we have at least three inputs. After numerous attempts, we determine the 3-6-1 feedforward backpropagation neural network as prediction model (**Figure 1**), *i.e.* the first layer contains three neurons corresponding to three inputs (or three elements of input in neural network terminology), the second layer contains six neurons, and the last layer contains one neuron corresponding to the target. The transfer functions for three layers are tan-sigmoid, tan-sigmoid and log-sigmoid, respectively. The training algorithm is the resilient backpropagation, which is the fastest algorithm on pattern recognition [18]. This is because the resilient backpropagation uses the sign of the derivative rather than the magnitude of the derivative to determine the direction of convergence [26].

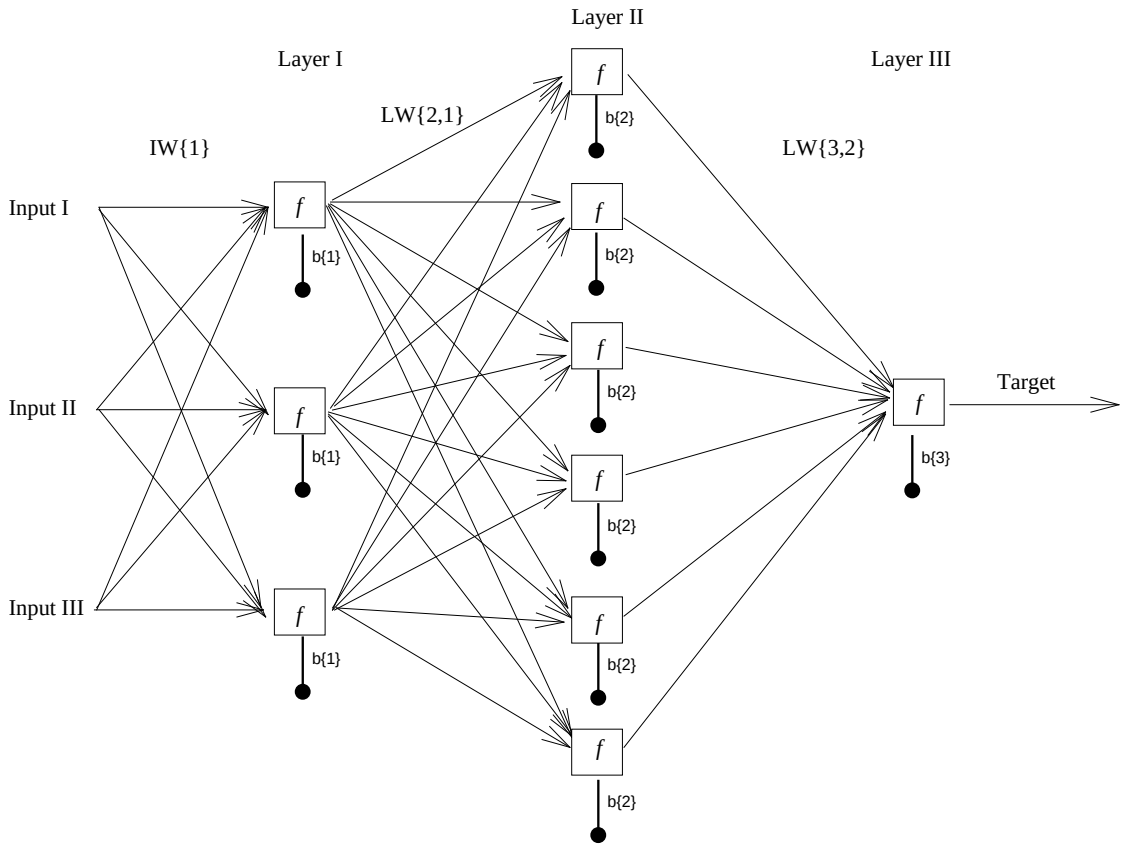

Figure 1. The 3-6-1 feedforward backpropagation neural network. Each square presents a neuron. IW{1} is the input weights, LW{2, 1} is the layer weights to the second layer from the first layer, and LW{3, 2} is the layer weights to the third layer from the second layer. b{1}, b{2} and b{3} are the biases related to each neuron at the first, second and third layers, respectively. *f* is the transfer function.

Input I—Amino-acid pair predictability

We calculate this random quantification according to the permutation, and we have used it in studying various proteins [12-15]. In general, this quantification is very sensitive to the change in neighbouring amino acids, and answers why a type of amino acid is adjacent to a certain type of amino acid but not to the others. Besides, the reason for using amino-acid pair is that a good signature pattern of a protein must be as short as possible, but the conserved sequence is not longer than four or five residues [27], and we had used three-amino-acid, four-amino-acid, and five-amino acid sequences in our earliest studies, and found the amino-acid pair best suitable.

The simplest calculations are as follows. According to permutation, for example, there are 52 glycines (G) and 45 isoleucines (I) in the hemagglutinin, strain A/mallard/Sweden/85/02(H7N7), accession number AY999979, the frequency of amino-acid pair "GI" is 4 ($51/560 \times 45/559 \times 559 = 4.0982$), that is, the "GI" would appear four times in this hemagglutinin. Actually we do find 4 "GI", so the amino-acid pair "GI" is predictable and the difference between its actual and predicted frequencies is 0. Again, there are 27 aspartic acids (D) in AY999979 hemagglutinin, and the frequency of random presence of "ID" is 2 ($45/560 \times 27/559 \times 559 = 2.1696$), *i.e.* there would be two "ID" in the hemagglutinin. But the "ID" appears eight times in reality, so the difference between its actual and predicted frequency is 6. After such calculations, each amino-acid pair has its difference between actual and predicted frequencies. As a point mutation is relevant to a single amino acid, which connects with two neighbouring amino acids except for the terminal one and constructs two amino-acid pairs, so each amino acid has the sum of difference between actual and predicted frequencies in two neighbouring amino-acid pairs.

Input II—Amino-acid distribution probability

We calculate this random quantification according to the occupancy of subpopulations and partitions [28], and we have used it [12-15]. In general, this quantification is mainly subject to any change in the position of amino acid, and answers why the majority of amino acids cluster in some regions rather than homogenously distribute along the primary structure of a protein.

The quantification is developed along such thought, for example, there are two methionines (M) among 141 amino acids in human hemoglobin α-chain. With regard to their random distribution, our intuition may suggest that there would be one M in the first half of the chain and another M in the second half, which is true in real-life case. In fact, there are only three possible distributions of Ms in human hemoglobin α-chain, *i.e.* 1) both Ms are in the first half, 2) one M is in each half and 3) both Ms are in the second half. If we do not distinguish either first half or second half but are simply interested in whether both Ms are in both halves or in any half, we will have the probability of 1/2 for each distribution.

If we are interested in the distribution probability of three amino acids in a protein, we naturally imagine to group the protein into three parts, and our intuition may suggest that each part contains an amino acid. If we do not distinguish the first, second and third part, actually there are total three types of distributions, *i.e.* 1) three amino acids are in each part, 2) two amino acids are in a part and an amino acid in another part, and 3) three amino acids are in a part. However, the distribution probabilities are different for these three types of distributions, say, 0.2222 for 1), 0.6667 for 2) and 0.1111 for 3). Clearly the protein can only adopt one type of distribution for these three amino acids, which is the actual distribution probability, and we may guess that the distribution 2) is more likely to happen because of its highest probability, which is the predicted distribution probability.

For four amino acids, we will have five distribution probabilities, *i.e.* 1) each part contains an amino acid, 2) a part contains two amino acids and two parts contain an amino acid each, 3) two parts contain two amino acids each, 4) a part contains an amino acid and a part contains three amino acids, and 5) a part contains four amino acids. Their distribution probabilities are 0.0938 for 1), 0.5625 for 2), 0.1406 for 3), 0.1875 for 4) and 0.0156 for 5). Further, we have seven distributions for five amino acids, we have 11 distributions for six amino acids, we have 15 distributions for seven amino acids, and so on.

So we view the positions of each kind of amino acids in a protein as a distribution, whose probability can be calculated according to the equation of $r!/(q0! \times q1! \times \ldots \times qn!) \times r!/(r1! \times r2! \times \ldots \times rn!) \times n^{-r}$, where! is the factorial function, r is the number of a kind of amino acid, q is the number of parts with the same number of amino acids and n is the number of grouped parts in the protein for a kind of amino acid.

For instance, there are 38 leucines (L) in AY999979 hemagglutinin, whose predicted and actual distribution probabilities are 0.0373 and 0.0071, so the ratio of predicted versus actual distribution probabilities is 5.2535, whose natural logarithm is 1.6589, which is assigned to each L in AY999979 hemagglutinin.

In fact, this distribution probability can be referred to the statistical mechanics, which classifies the distribution of elementary particles in energy states according to three assumptions of whether or not distinguishing of each particle and energy state, *i.e.* Maxwell-Boltzmann, Fermi-Dirac and Bose-Einstein assumptions [28]. In plain words, this distribution probability is the probability if we would receive seven letters in a week but the letters distribute randomly.

Input III—Future composition of amino acids

We calculate this random quantification according to the translation probability between RNA codons and translated amino acids [29]. In general, this quantification is mainly subject to the future mutation trend, and answers what probability an amino acid mutates to another type of amino acid.

This quantification is developed along such line of thought, for example, we are interested in the amino acid "L" and its mutated amino acids with their mutating probability. As the RNA codons have the unambiguous relationship with their translated amino acids, we can extend this question to RNA level, this is, a point mutation in RNA codon leads to the mutation at amino acid level.

The "L" is related to RNA codons UUA, UUG, CUU, CUC, CUA and CUG, the mutation at the first position of UUA can lead UUA to mutate to AUA, CUA and GUA, which correspond to "L" to mutate to "I", "L", and "V' at amino acid level. Similarly, the mutation at second position of UUA results in "STOP", "STOP", and "S" at amino acid level, the mutation at the third position of UUA results in "F", "L" and "F" at amino acid level. Taken six RNA codons together, "L" would mutate in such a way, say, 6F + 2H + 4I + 18L + 2M + 4P + 2Q + 4R + 2S + 1W + 6V + 3STOP. Thus we have the L mutating probability to these amino acids, say, 6/54 + 2/54 + 4/54 + 18/54 + 2/54 + 4/54 + 2/54 + 4/54 + 2/54 + 1/54 + 6/54 + 3/54. For all 20 kinds of amino acids, we have the amino acid mutating probability in **Table 1**.

For the calculation of future composition of amino acids, we have the following steps: 1) we would expect that "A" has the 12/36 chance of mutating to "A" (line 2 in **Table 1**), "R" and "N" have no chance of mutating to "A" (lines 3 and 4 in **Table 1**), "D" has 2/18 chance (line 5 in **Table 1**), "C" has no chance (line 6 in **Table 1**), "E" has 2/18 chance, and so on. 2) Meanwhile, AY999979 hemagglutinin has 35 "A", 30 "R", 39 "N", 27 "D", 16 "C", 35 "E", and so on. 3) So we can estimate how many "A" can be mutated using $35 \times 12/36 + 30 \times 0 + 39 \times 0 + 27 \times 2/18 + 16 \times 0 + 35 \times 2/18$, and so on. In total, this is the future composition of amino acid "A". 4) After calculated all 20 kinds of amino acids, "A" contributes 6.4352% of future composition in hemagglutinin. 5) On the other hand, "A" contributes 6.25% (35/560) of current composition in AY999979 hemagglutinin. 6) Thus, we have the ratio of future versus current compositions, for example, the ratio of "A" is 1.0296 (6.4352%/6.25%), which can be assigned to each "A" in AY999979 hemagglutinin. 7) In this manner, we have the future compositions for all amino acids.

Target—Occurrence or non-occurrence of mutation

The phylogenetics analyses the evolutionary process of hemagglutinins in question. Along same branch of the evolutionary tree, we can compare the parent and daughter hemagglutinins, the difference between them indicates the occurrence of mutation, which we mark as unity, whereas no difference between them indicates the non-occurrence of mutation, which we mark as zero.

Prediction of would-be-mutated amino acids

Currently, we have no explicit idea to build a cause-mutation relationship between an original amino acid and its mutated amino acids. However, we can make the estimation according to the amino acid mutating probability in **Table 1**. For instance, if we predict that the mutation position is 237, which houses amino acid "L", from **Table 1** we know that "L" has the largest chance of mutating to "F" and "V" if we do not consider the case that "L" mutates to "L", and then the equal chance of mutating to "I", "P" and "R", and so on. In this manner, we make the prediction.

Software and statistics

The MatLab software [30] is used for the model development and prediction. The outlier (3SD) is detected according to the published method [31]. The calculations of prediction sensitivity, specificity and total correct rate are according to the published method [32].

Table 1. Amino acid mutating probability based on the translation probability between RNA codons and translated amino acids.

Amino acid	Mutated amino acid with its translation probability
A	12/36A + 2/36D + 2/36E + 4/36G + 4/36P + 4/36S + 4/36T + 4/36V
R	2/54C + 6/54G + 2/54H + 1/54I + 2/54K + 4/54L + 1/54M + 4/54P + 2/54Q + 18/54R + 6/54S + 2/54T + 2/54W + 2/54STOP
N	2/18D + 2/18H + 2/18I + 4/18K + 2/18N + 2/18S + 2/18T + 2/18Y
D	2/18A + 2/18D + 4/18E + 2/18G + 2/18H + 2/18N + 2/18V + 2/18Y
C	2/18C + 2/18F + 2/18G + 2/18R + 4/18S + 2/18W + 2/18Y + 2/18STOP
E	2/18A + 4/18D + 2/18E + 2/18G + 2/18K + 2/18Q + 2/18V + 2/18STOP
Q	2/18E + 4/18H + 2/18K + 2/18L + 2/18P + 2/18Q + 2/18R + 2/18STOP
G	4/36A + 2/36C + 2/36D + 2/36E + 12/36G + 6/36R + 2/36S + 4/36V + 1/36W + 1/36STOP
H	2/18D + 2/18H + 2/18L + 2/18N + 2/18P + 4/18Q + 2/18R + 2/18Y
I	2/27F + 6/27I + 1/27K + 4/27L + 3/27M + 2/27N + 1/27R + 2/27S + 3/27T + 3/27V
L	6/54F + 2/54H + 4/54I + 18/54L + 2/54M + 4/54P + 2/54Q + 4/54R + 2/54S + 1/54W + 6/54V + 3/54STOP
K	2/18E + 1/18I + 2/18K + 1/18M + 4/18N + 2/18Q + 2/18R + 2/18T + 2/18STOP
M	3/9I + 1/9K + 2/9L + 1/9R + 1/9T + 1/9V
F	2/18C + 2/18F + 2/18I + 6/18L + 2/18S + 2/18V + 2/18Y
P	4/36A + 2/36H + 4/36L + 12/36P + 2/36Q + 4/36R + 4/36S + 4/36T
S	4/54A + 4/54C + 2/54F + 2/54G + 2/54I + 2/54L + 2/54N + 4/54P + 6/54R + 14/54S + 6/54T + 1/54W + 2/54Y + 3/54STOP
T	4/36A + 3/36I + 2/36K + 1/36M + 2/36N + 4/36P + 2/36R + 6/36S + 12/36T
W	2/9C + 1/9G + 1/9L + 2/9R + 1/9S + 2/9STOP
Y	2/18C + 2/18D + 2/18F + 2/18H + 2/18N + 2/18S + 2/18Y + 4/18STOP
V	4/36A + 2/36D + 2/36E + 2/36F + 4/36G + 3/36I + 6/36L + 1/36M + 12/36V
STOP	1/27C + 2/27E + 1/27G + 2/27K + 3/27L + 2/27Q + 2/27R + 3/27S + 2/27W + 4/27Y + 4/27STOP

A, alanine; R, arginine; N, asparagine; D, aspartic acid; C, cysteine; E, glutamic acid; Q, glutamine; G, glycine; H, histidine; I, isoleucine; L, leucine; K, lysine; M, methionine; F, phenylalanine; P, proline; S, serine; T, threonine; W, tryptophan; Y, tyrosine; V, valine.

3. RESULTS AND DISCUSSION

The quantification of parent hemagglutinins leads each parent hemagglutinin to have three numerical inputs and a binary target in each amino acid, between them is the hidden cause-mutation relationship, which we hope to use the neural network to enhance the predictability. Table 2 shows a fraction of a parent hemagglutinin after comparing two swine hemagglutinins (AY999979 and AY999980). Thus, we can use this format of data to train the neural network.

After tried different neural network models with different numbers of layers, neurons, transfer functions, training algorithms, we determine the 3-6-1 feedforward backpropagation neural network as suitable model, the tan-sigmoid, tan-sigmoid and log-sigmoid as suitable transfer functions and the resilient backpropagation as suitable training algorithm (Figure 1). As no historical data on the initial weights and biases are available for our neural network, we use the random initialisation function to initiate the neural network. Although we can raise the question of whether the neural network can converge during its training with a limited number of epochs, the preliminary training shows that the neural network converge within 250 epochs, even the initial weights and biases were randomly given by the initialisation function. Hence, we can use the random initialisation function to train the neural network to find the suitable weights and biases.

Yet, we need to determine whether the neural network can capture the cause-mutation relationship defined by internal randomness. We can classify the predicted mutation positions as the positives, false positives, negatives and false negatives when comparing the predicted with the actual mutation positions, and then use the prediction sensitivity, specificity and total correct rate (Figure 2) to evaluate the prediction performance. As seen, the prediction specificity and total correct rate are quite similar between the prediction made by neural network and the prediction made by logistic regression, the prediction sensitivity in far better in the prediction made by neural network than the prediction made by logistic regression. This means that the neural network indeed enhanced the predictability.

After prediction of mutation positions by the neural network, we can predict the would-be-mutated amino acids at predicted positions with the help of amino acid mutating probability in Table 1. Figure 3 illustrates the two-step frame in prediction of mutation. The solid line in the lower panel is the predicted mutation probability by the neural network with respect to DQ873807 hemagglutinin and the dash-dotted line is the cut-off mutation probability of 0.5, beyond which the amino acid risks mutation.

Table 2. Inputs and target of AY999979 hemagglutinin.

Position	Amino acid	Input			Target
		I	II	III	
1	M	2	0.0000	0.7315	0
…	…	…	…	…	…
164	N	4	2.4350	0.5845	0
165	T	2	0.5596	1.0476	1
166	D	−1	0.5978	0.8992	0
167	N	5	2.4350	0.5845	0
168	A	5	0.1178	1.0296	1
…	…	…	…	…	…
560	I	1	0.3060	0.7420	0

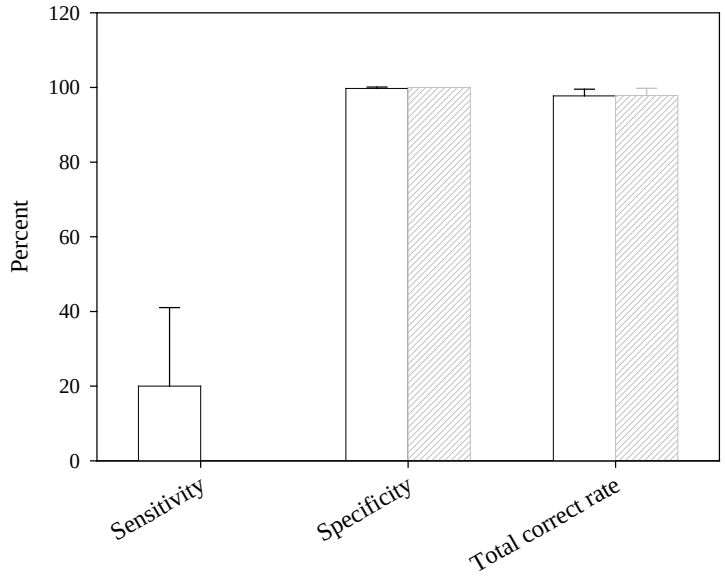

Figure 2. Prediction sensitivity, specificity and total correct rate between the prediction made by neural network (unshaded bars) and the prediction made by logistic regression (shaded bars). The data are presented as mean ± SD (n = 84). The sensitivity is equal to the predicted positives/the actual mutations (%), the specificity is equal to the predicted negatives/the actual non-mutations (%), and the total correct rate is equal to (predicted positives + predicted negatives)/length of hemagglutinin (%).

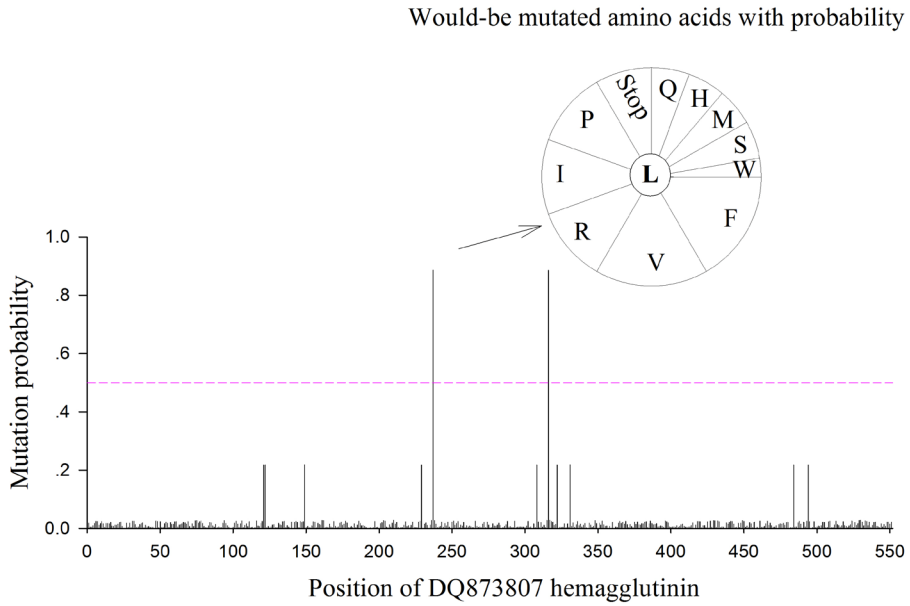

Figure 3. Prediction of mutation in DQ873807 hemagglutinin. The neural network used the following weights and biases: IW{1, 1} = [−0.2831 1.2534 −2.9981; −0.0928 −0.0935 −1.4680; 0.2819 0.9855 2.3676], LW{2, 1} = [4.3791 −6.0042 −2.1001; 2.3406 1.6956 2.6433; 46.7964 0.0119 −17.0226; 46.8219 77.5225 746.4238; 2.6449 −5.6936 −2.2864; 0.9433 −1.2682 −0.6735], LW{3, 2} = [−1.6174 −5.6434 1.6643 4.4668 −1.4110 −2.7877], b{1} = [4.9548 1.4145 −5.4458], b{2} = [20.9094 −1.1345 1.0613 −14.1254 17.3764 18.4171], b{3} = [−0.5444].

The predictions in Figure 3 are as follows: 1) there are potential 11 mutations as shown 11 solid vertical lines on *x*-axis, but only potential 2 mutations have a mutation probability larger than 0.5 while others have the mutation probability equal to 0.2; 2) there are two positions whose mutation probability is larger than 0.5 are positions 237 and 316, and the amino acids at these two positions are "L".

If these two "L" at positions 237 and 316 would mutate according to the model prediction, what types of amino acids will they mutate to? For this question, we refer to Table 1, which probabilistically tells what amino acids will be mutated.

So "L" at would have a larger chance of occurring of mutation. The upper panel gives the estimation of would-be-mutated amino acid at the predicted position according to the amino acid mutating probability in Table 1. When looking at the line initiated with "L", we can find that amino acid "L" has the following probabilities to mutate to other amino acids, *i.e.* 6/54F + 2/54H + 4/54I + 18/54L + 2/54M + 4/54P + 2/54Q + 4/54R + 2/54S + 1/54W + 6/54V + 3/54STOP, which depicted as a pie according to their portions. Accordingly, "L" has the great chance to mutate to "F" (6/54 chance).

In this way, we can predict the potential mutation positions with mutated amino acid for each protein sequence. But we must admit several limitations, which require for improvements in the future. 1) This prediction is based on the probabilistic model, so the maximal probability suggests the occurrence of mutation, but mutation may randomly occur at a position with low probability and mutate to an amino acid with low probability. This requires us to examine the database to have an overall concept on the percentages of how many mutations occur with the maximal probability as well as other probabilities and how many amino acids mutate to amino acids with maximal probability as well as other probabilities. 2) The current database in fact does not indicate which sequence comes from which sequence, so it is very hard to verify the limitation in point 1, and verify our predictions.

Currently, we have yet to know how many types of internal power we can define, because nature is not designed according to our definitions; however randomness deems a representative power of nature. Among defined internal power, more complicated and difficult is how many types of internal power we can quantify.

Yet, randomness suggests that the construction of a protein requires the least time and energy, although nature would deliberately spend more time and energy to construct the absolutely necessary structure. This is identical to the parsimony in nature.

Nevertheless, there are a lot rooms for development of different models to predict the mutations in DNA/RNA/proteins because these prediction will reveal how they evolved in the past. There are more rooms for us to improve our predictive methods. Both require more work in the future.

In the past, not many studies were conducted for predictions of different aspects of H7 hemagglutinins, for example, the prediction of conserved B-cell epitopes of hemagglutinin H7 [33]. In fact, this type of predictions is to predict the conserved amino acid sequences, *i.e.* a amino acid pattern, whereas our prediction is not concerning with conserved amino acids but mutated amino acids.

4. CONCLUSION

In this study, we show the possibility to predict the mutation in H7 hemagglutinins from influenza A virus. Actually, the prediction is dynamic because it changes according to the given amino acid sequences and is weighted by phylogenetic relationship between father-daughter mutation patterns. Nevertheless, more studies are needed to develop and valid the current method in near future.

FUND

This study was partly supported by National Natural Science Foundation of China (31460296 and 31560315), Key Project of Guangxi Scientific Research and Technology Development Plan (AB17190534).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest regarding the publication of this paper.

REFERENCES

1. Martini, M., Gazzaniga, V., Bragazzi, N.L. and Barberis, I. (2019) The Spanish Influenza Pandemic: A Lesson from History 100 Years after 1918. *Journal of Preventive Medicine and Hygiene*, **60**, E64-E67.
2. Qin, Y., Zhao, M.J., Tan, Y.Y., Li, X.Q., Zheng, J.D., Peng, Z.B. and Feng, L.Z. (2018) History of Influenza Pandemics in China during the Past Century. *Chinese Journal of Epidemiology*, **39**, 1028-1031.
3. Belser, J.A., Bridges, C.B., Katz, J.M. and Tumpey, T.M. (2009) Past, Present, and Possible Future Human Infection with Influenza Virus A Subtype H7. *Emerging Infectious Diseases*, **15**, 59-65. https://doi.org/10.3201/eid1506.090072
4. Fouchier, R.A., Schneeberger, P.M., Rozendaal, F.W., Broekman, J.M., Kemink, S.A., Munster, V., Kuiken, T., Rimmelzwaan, G.F., Schutten, M., Van Doornum, G.J., Koch, G., Bosman, A., Koopmans, M. and Osterhaus, A.D. (2004) Avian Influenza A Virus (H7N7) Associated with Human Conjunctivitis and a Fatal Case of Acute Respiratory Distress Syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 1356-1361. https://doi.org/10.1073/pnas.0308352100
5. Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., Chen, J., Jie, Z., Qiu, H., Xu, K., Xu, X., Lu, H., Zhu, W., Gao, Z., Xiang, N., Shen, Y., He, Z., Gu, Y., Zhang, Z., Yang, Y., Zhao, X., Zhou, L., Li, X., Zou, S., Zhang, Y., Li, X., Yang, L., Guo, J., Dong, J., Li, Q., Dong, L., Zhu, Y., Bai, T., Wang, S., Hao, P., Yang, W., Zhang, Y., Han, J., Yu, H., Li, D., Gao, G.F., Wu, G., Wang, Y., Yuan, Z. and Shu, Y. (2013) Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus. *New England Journal of Medicine*, **368**, 1888-1897. https://doi.org/10.1056/NEJMoa1304459
6. Li, Q., Zhou, L., Zhou, M., Chen, Z., Li, F., Wu, H., Xiang, N., Chen, E., Tang, F., Wang, D., Meng, L., Hong, Z., Tu, W., Cao, Y., Li, L., Ding, F., Liu, B., Wang, M., Xie, R., Gao, R., Li, X., Bai, T., Zou, S., He, J., Hu, J., Xu, Y., Chai, C., Wang, S., Gao, Y., Jin, L., Zhang, Y., Luo, H., Yu, H., He, J., Li, Q., Wang, X., Gao, L., Pang, X., Liu, G., Yan, Y., Yuan, H., Shu, Y., Yang, W., Wang, Y., Wu, F., Uyeki, T.M. and Feng, Z. (2014) Epidemiology of Human Infections with Avian Influenza A(H7N9) Virus in China. *New England Journal of Medicine*, **370**, 520-532. https://doi.org/10.1056/NEJMoa1304617
7. Liu, D., Shi, W., Shi, Y., Wang, D., Xiao, H., Li, W., Bi, Y., Wu, Y., Li, X., Yan, J., Liu, W., Zhao, G., Yang, W., Wang, Y., Ma, J., Shu, Y., Lei, F. and Gao, G.F. (2013) Origin and Diversity of Novel Avian Influenza A H7N9 Viruses Causing Human Infection: Phylogenetic, Structural, and Coalescent Analyses. *The Lancet*, **381**, 1926-1932. https://doi.org/10.1016/S0140-6736(13)60938-1
8. Zhou, L., Ren, R., Yang, L., Bao, C., Wu, J., Wang, D., Li, C., Xiang, N., Wang, Y., Li, D., Sui, H., Shu, Y., Feng, Z., Li, Q. and Ni, D. (2017) Sudden Increase in Human Infection with Avian Influenza A(H7N9) Virus in China, September-December 2016. *Western Pacific Surveillance and Response Journal*, **8**, 6-14. https://doi.org/10.5365/wpsar.2017.8.1.001
9. Fadlallah, G.M., Ma, F., Zhang, Z., Hao, M., Hu, J., Li, M., Liu, H., Liang, B., Yao, Y., Gong, R., Zhang, B., Liu, D. and Chen, J. (2020) Vaccination with Consensus H7 Elicits Broadly Reactive and Protective Antibodies against Eurasian and North American Lineage H7 Viruses. *Vaccines* (*Basel*), **8**, 143. https://doi.org/10.3390/vaccines8010143
10. Qi, W., Jia, W., Liu, D., Li, J., Bi, Y., Xie, S., Li, B., Hu, T., Du, Y., Xing, L., Zhang, J., Zhang, F., Wei, X., Eden, J.S., Li, H., Tian, H., Li, W., Su, G., Lao, G., Xu, C., Xu, B., Liu, W., Zhang, G., Ren, T., Holmes, E.C., Cui, J., Shi, W., Gao, G.F. and Lia, M. (2018) Emergence and Adaptation of a Novel Highly Pathogenic H7N9 Influenza Virus in Birds and Humans from a 2013 Human-Infecting Low-Pathogenic Ancestor. *Journal of Virology*, **92**, e00921-17. https://doi.org/10.1128/JVI.00921-17
11. Everitt, B.S. (1999) Chance Rules: An Informal Guide to Probability, Risk, and Statistics. Springer, New York.
12. Wu, G. and Yan, S. (2002) Randomness in the Primary Structure of Protein: Methods and Implications. *Mo-*

lecular Biology Today, **3**, 55-69. https://doi.org/10.1063/1.2408446

13. Wu, G. and Yan, S. (2006) Mutation Trend of Hemagglutinin of Influenza A Virus: A Review from Computational Mutation Viewpoint. *Acta Pharmacologia Sinica*, **27**, 513-526. https://doi.org/10.1111/j.1745-7254.2006.00329.x
14. Yan, S. and Wu, G. (2013) Predictions of Enzymatic Parameters: A Mini-Review with Focus on Enzymes for Biofuel. *Applied Biochemistry and Biotechnology Part A*: *Enzyme Engineering and Biotechnology*, **171**, 590-615. https://doi.org/10.1007/s12010-013-0328-6
15. Wu, G. and Yan, S. (2006) Fate of Influenza A Virus Proteins. *Protein & Peptide Letters*, **13**, 377-384. https://doi.org/10.2174/092986606775974474
16. Draper, N.R. and Smith, H. (1981) Applied Regression Analysis. 2nd Edition, Wiley, New York.
17. Hosmer Jr., D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd Edition, Wiley, New York. https://doi.org/10.1002/0471722146
18. Demuth, H. and Beale, M. (2001) Neural Network Toolbox for Use with MatLab. User's Guide, Version 4.
19. Wiley, D.C. and Skehel, J.J. (1987) The Structure and Function of the Hemagglutinin Membrane Glycoprotein of Influenza Virus. *Annual Review of Biochemistry*, **56**, 365-394. https://doi.org/10.1146/annurev.bi.56.070187.002053
20. Russell, R.J., Gamblin, S.J., Haire, L.F., Stevens, D.J., Xiao, B., Ha, Y. and Skehel, J.J. (2004) H1 and H7 Influenza Haemagglutinin Structures Extend a Structural Classification of Haemagglutinin Subtypes. *Virology*, **325**, 287-296. https://doi.org/10.1016/j.virol.2004.04.040
21. Kaleta, E.F. and Hönicke, A. (2004) Review of the Literature on Avian Influenza A Viruses in Pigeons and Experimental Studies on the Susceptibility of Domestic Pigeons to Influenza A Viruses of the Haemagglutinin Subtype H7. *Deutsche tierarztliche Wochenschrift*, **111**, 467-472.
22. Ilyushina, N.A., Govorkova, E.A., Russell, C.J., Hoffmann, E. and Webster, R.G. (2007) Contribution of H7 Haemagglutinin to Amantadine Resistance and Infectivity of Influenza Virus. *Journal of General Virology*, **88**, 1266-1274. https://doi.org/10.1099/vir.0.82256-0
23. Influenza Virus Resources (2020). https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi?go=database
24. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) Clustal W and Clustal X Version 2.0. *Bioinformatics*, **23**, 2947-2948. https://doi.org/10.1093/bioinformatics/btm404
25. Chou, K.C. (2011) Some Remarks on Protein Attribute Prediction and Pseudo Amino Acid Composition (50th Anniversary Year Review). *Journal of Theoretical Biology*, **273**, 236-247. https://doi.org/10.1016/j.jtbi.2010.12.024
26. Riedmiller, M. and Braun, H. (1993) A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. *Proceedings of the IEEE International Conference on Neural Networks*, Vol. 1, 586-591. https://doi.org/10.1109/ICNN.1993.298623
27. Prosite (2020) A Dictionary of Protein Sites and Patterns User Manual. https://prosite.expasy.org
28. Feller, W. (1968) An Introduction to Probability Theory and Its Applications. 3rd Edition, Vol. I, Wiley, New York.
29. Wu, G. and Yan, S. (2007) Translation Probability between RNA Codons and Translated Amino Acids, and Its Applications to Protein Mutations. In: Ostrovskiy, M.H., Ed., *Leading-Edge Messenger RNA Research Communications*, Chapter 3, Nova Science Publishers, New York, 47-65.
30. Mathworks Inc. (2001) MatLab—The Language of Technical Computing (Version 6.1.0.450, Release 12.1),

1984-2001.

31. Healy, M.J.R. (1979) Outliers in Clinical Chemistry Quality-Control Schemes. *Clinical Chemistry*, **25**, 675-677. https://doi.org/10.1093/clinchem/25.5.675
32. SYSTAT Software Inc. (2017) Systat 13 for Windows, Version 13.
33. Wang, X., Sun, Q., Ye, Z., Hua, Y., Shao, N., Du, Y., Zhang, Q. and Wan, C. (2016) Computational Approach for Predicting the Conserved B-Cell Epitopes of Hemagglutinin H7 Subtype Influenza Virus. *Experimental and Therapeutic Medicine*, **12**, 2439-2446. https://doi.org/10.3892/etm.2016.3636